Opportunities for Neuromorphic Computing Co-Processors

Katie Schuman

Department of Electrical Engineering and Computer Science, University of Tennessee

ScalAH22 Workshop, November 2022

Opportunities for Neuromorphic Computing Co-Processors

Katie Schuman
Department of Electrical Engineering and Computer Science

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

TENNLab

- Four PIs at UTK:
  - Dr. Ahmed Aziz (Devices)
  - Dr. Garrett Rose (Architectures and Devices)
  - Dr. Jim Plank (Software and Applications)
  - Dr. Katie Schuman (Algorithms and Applications)
- Affiliated faculty at:
  - SUNY Polytechnic
  - George Mason University
  - University of Mississippi
  - Florida International University
  - Oak Ridge National Laboratory
- Since 2015:
  - 12 Master's and 7 PhD graduates
  - Alumni at Tesla, Garmin, Intel, Cisco, Amazon, Micron, TI, Microsoft, Google, Facebook, SalesForce

TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

https://neuromorphic.eecs.utk.edu/

Why should you care about novel brain-inspired computer architectures?

Looming End of Moore's Law
(And the end of Dennard scaling)

+

Artificial Intelligence and Machine Learning

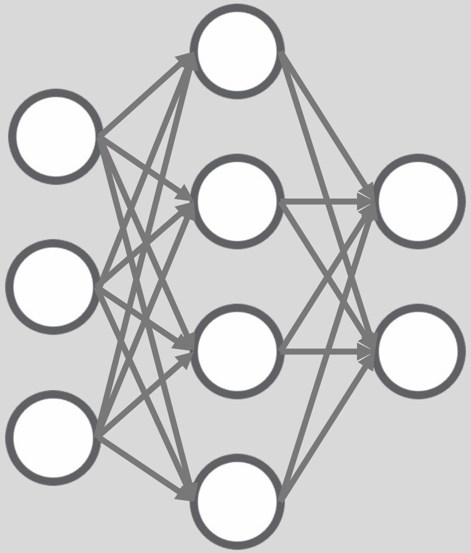
+

Rise of the Internet of Things

→ ?

Neural Hardware and Neuromorphic Computing

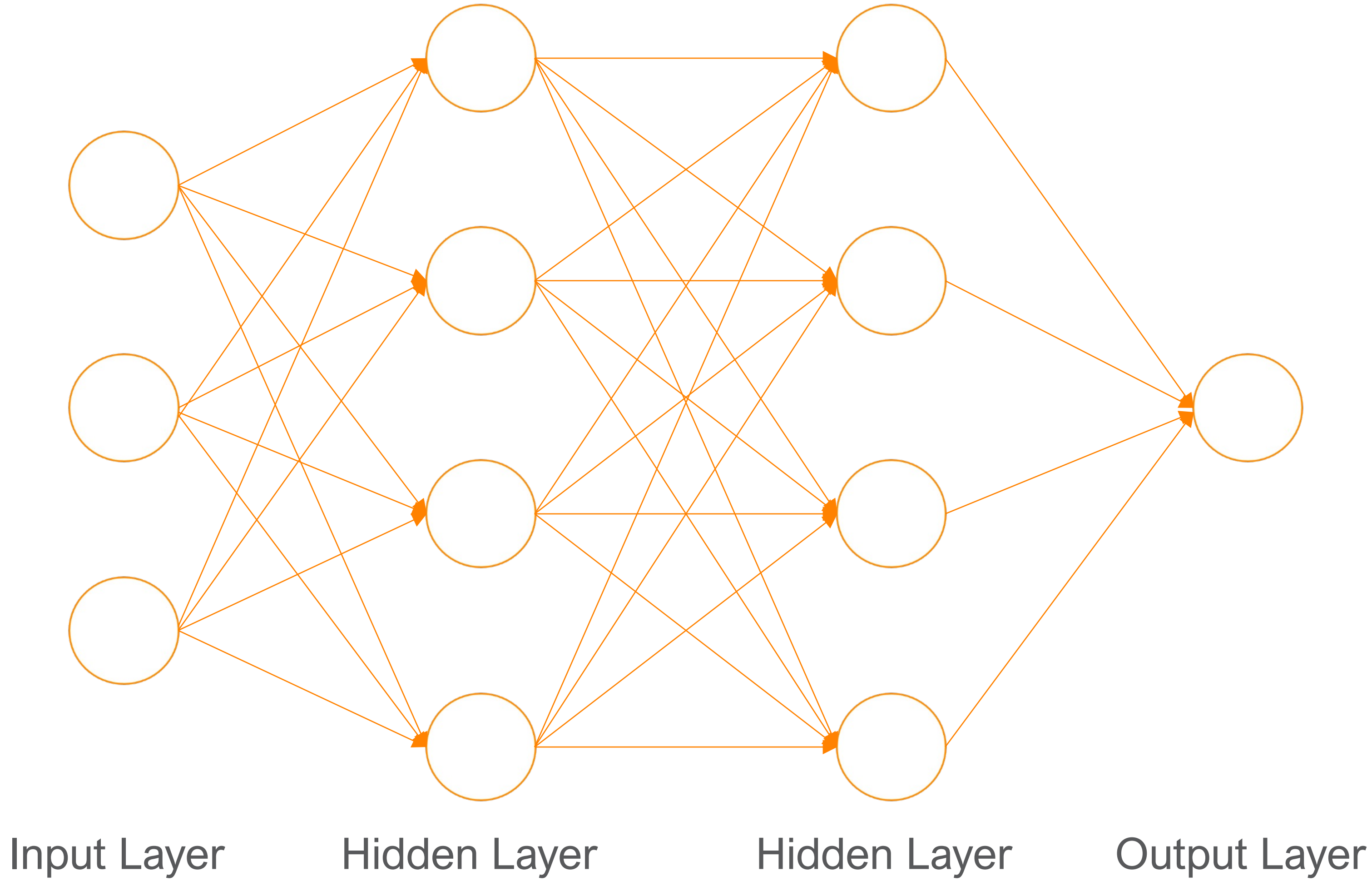
Neural Hardware



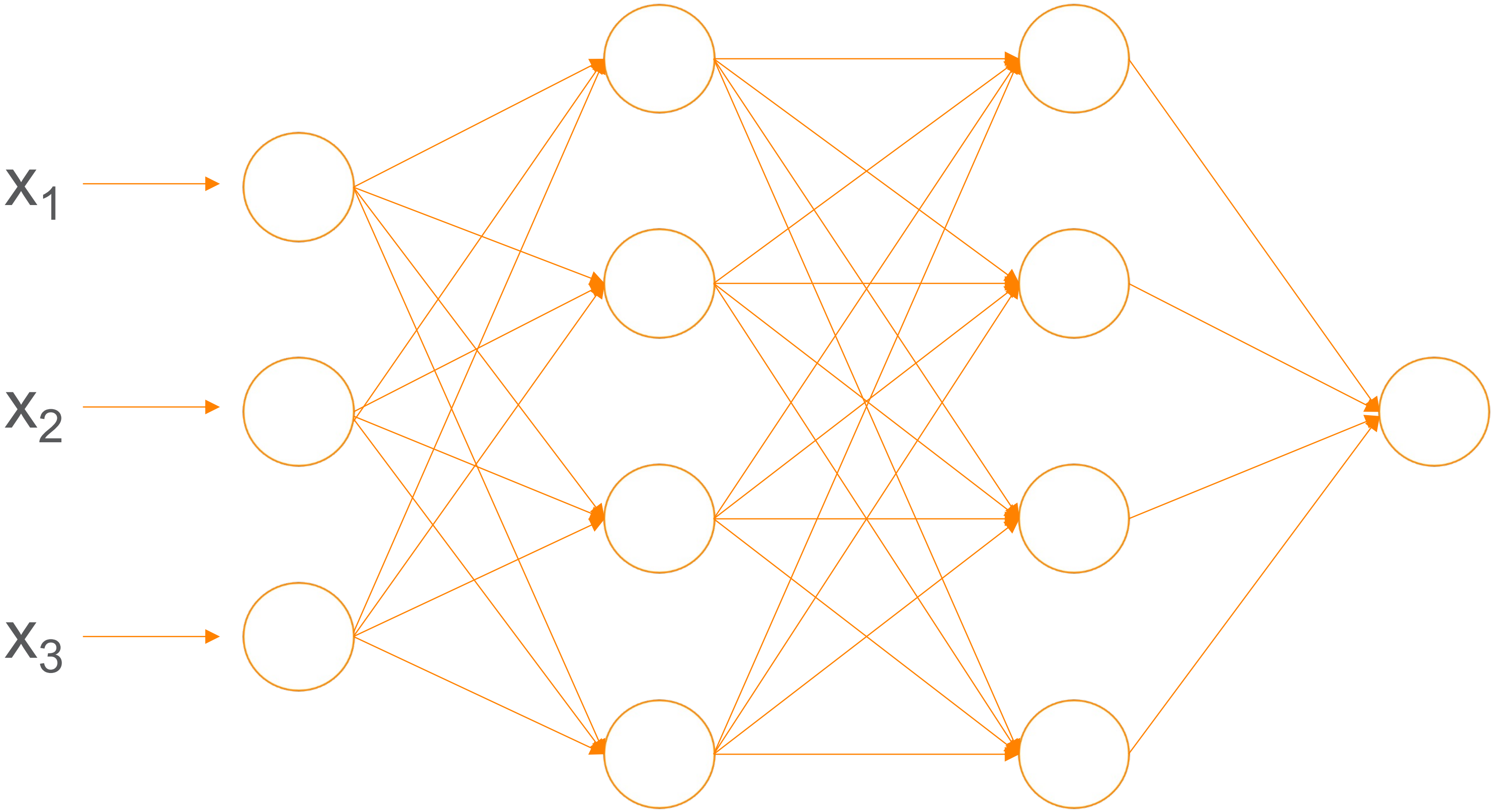
Accelerates traditional neural network and deep learning computation

- Well-suited to existing algorithms
- Fast computation ***or*** low power
- Currently deployed in cloud or mobile devices

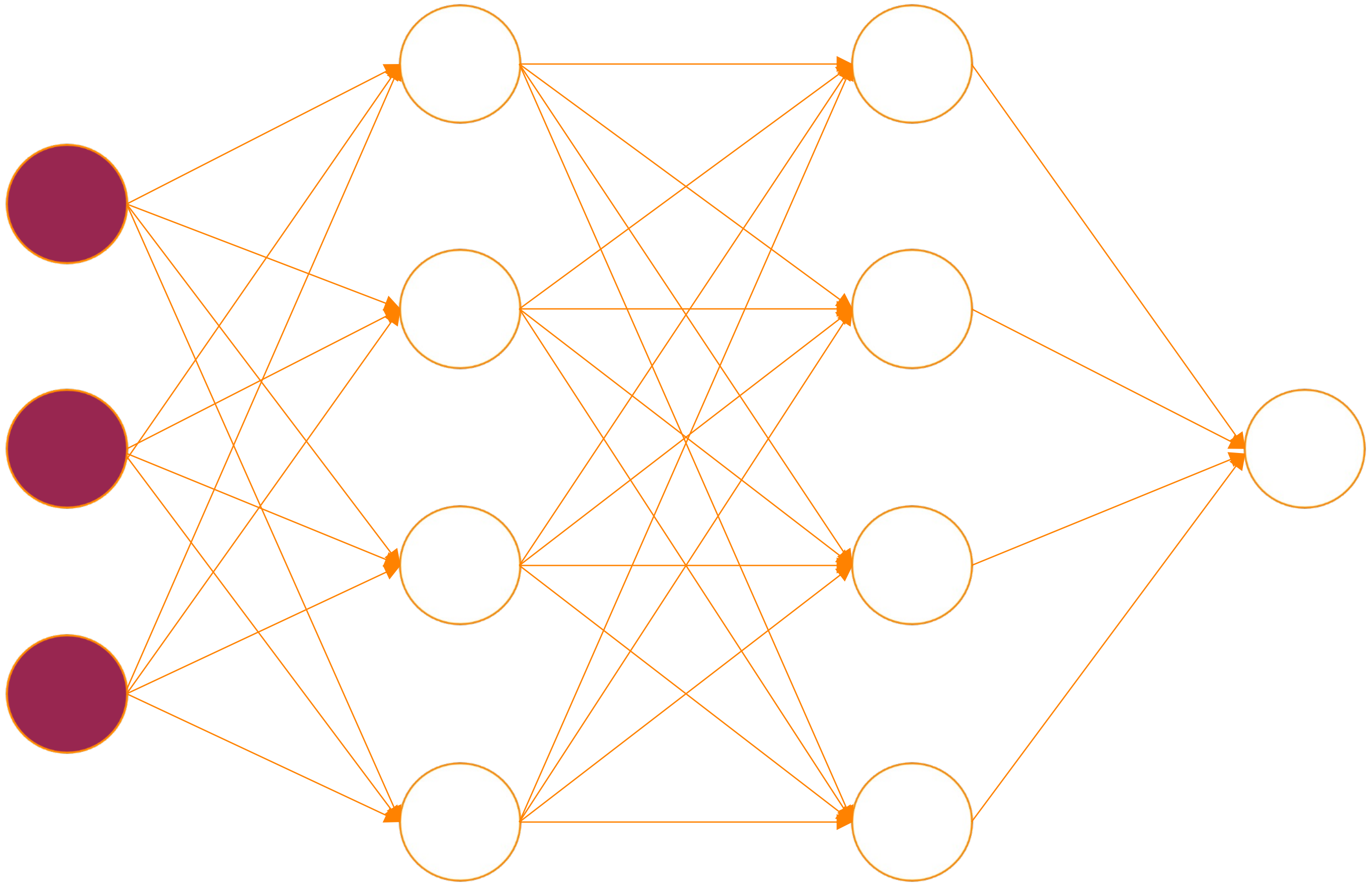
Traditional Artificial Neural Networks



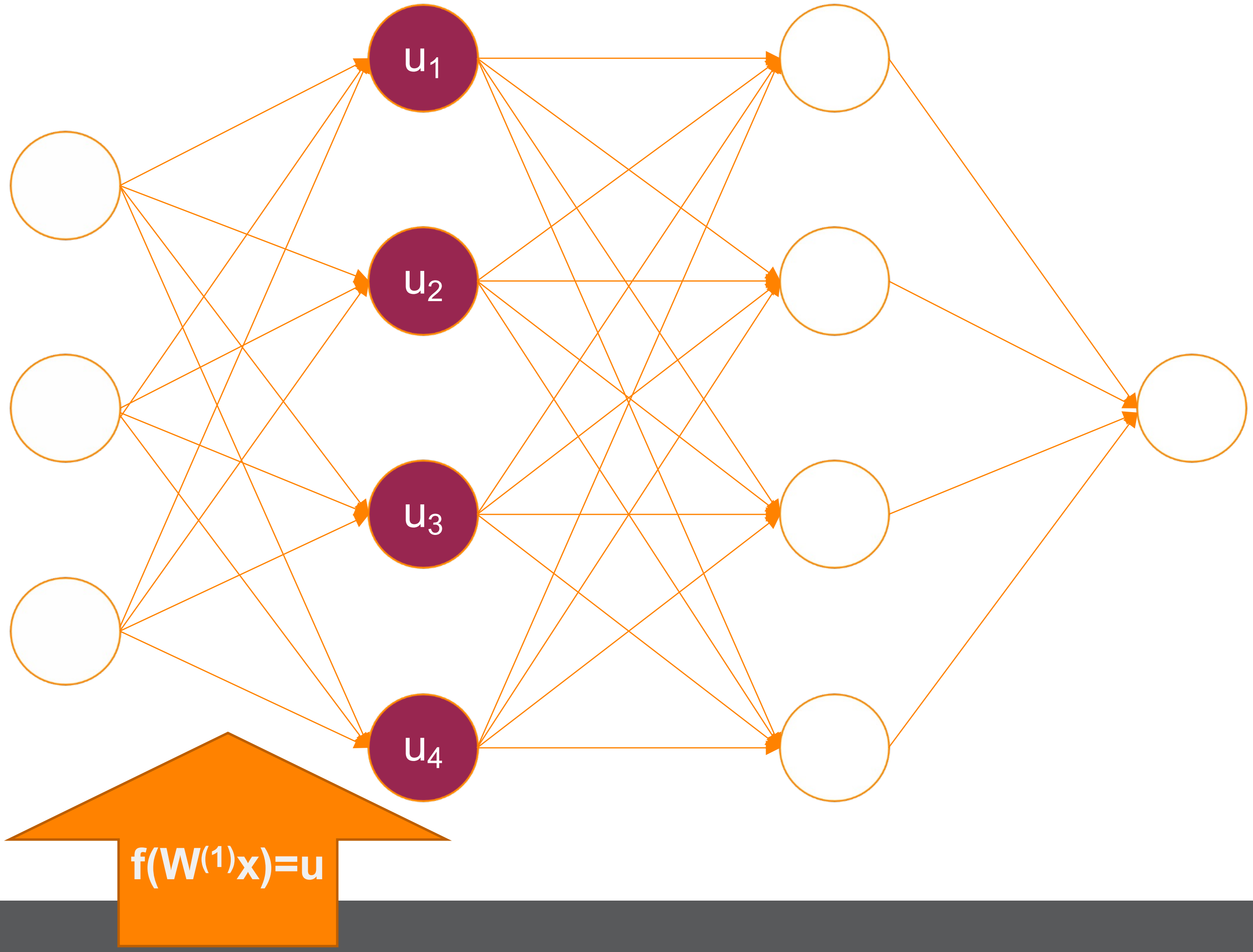
Traditional Artificial Neural Networks



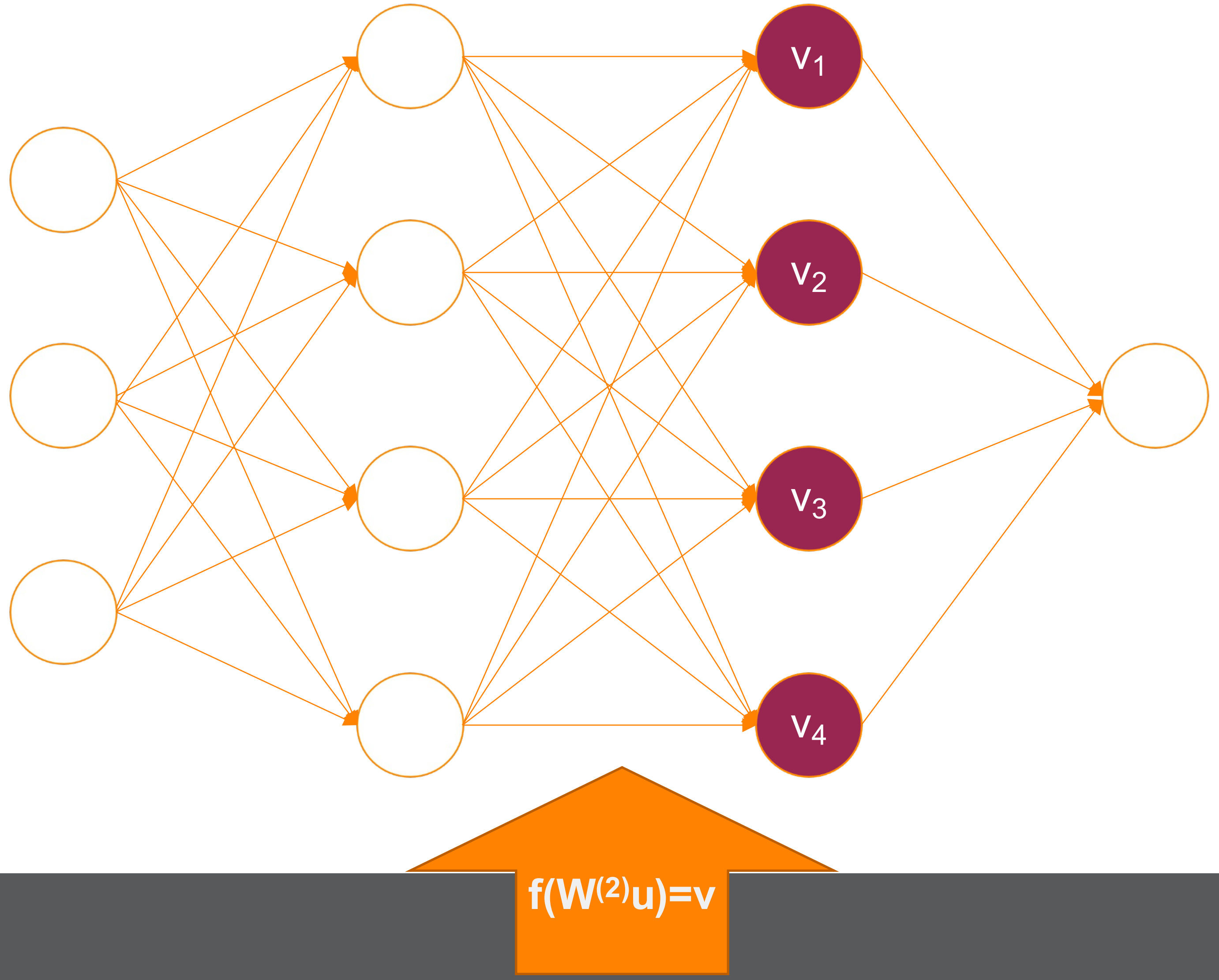
Traditional Artificial Neural Networks



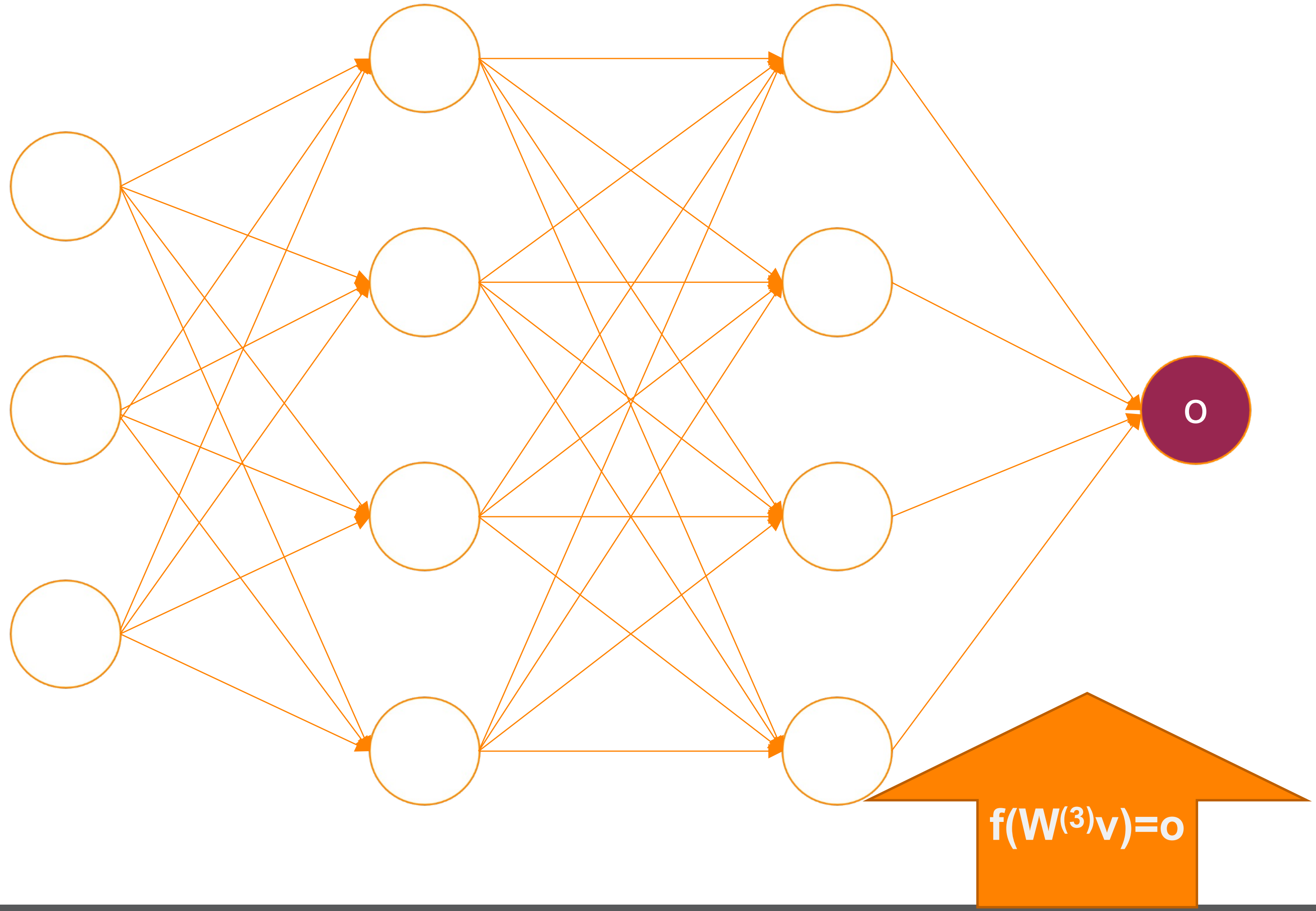
Traditional Artificial Neural Networks



Traditional Artificial Neural Networks

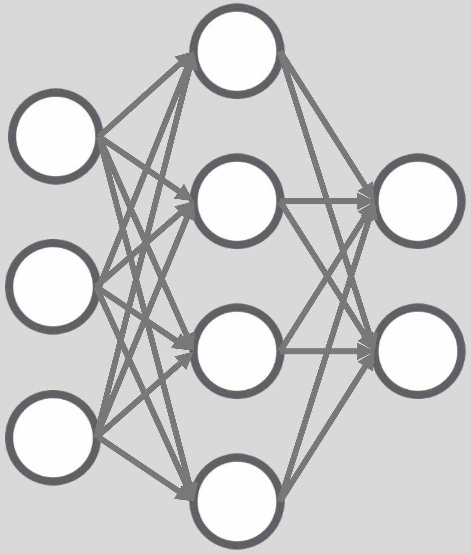


Traditional Artificial Neural Networks



Neural Hardware and Neuromorphic Computing

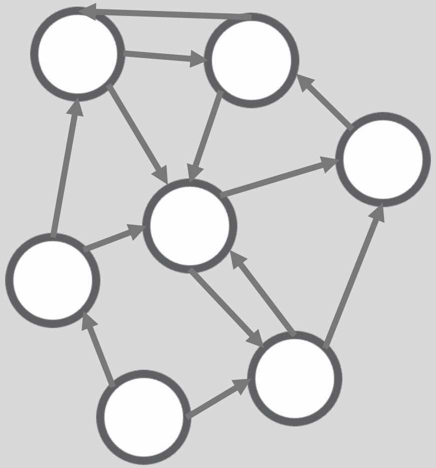
Neural Hardware



Accelerates traditional neural network and deep learning computation

- Well-suited to existing algorithms
- Fast computation ***or*** low power
- Currently deployed in cloud or mobile devices

Neuromorphic Computing

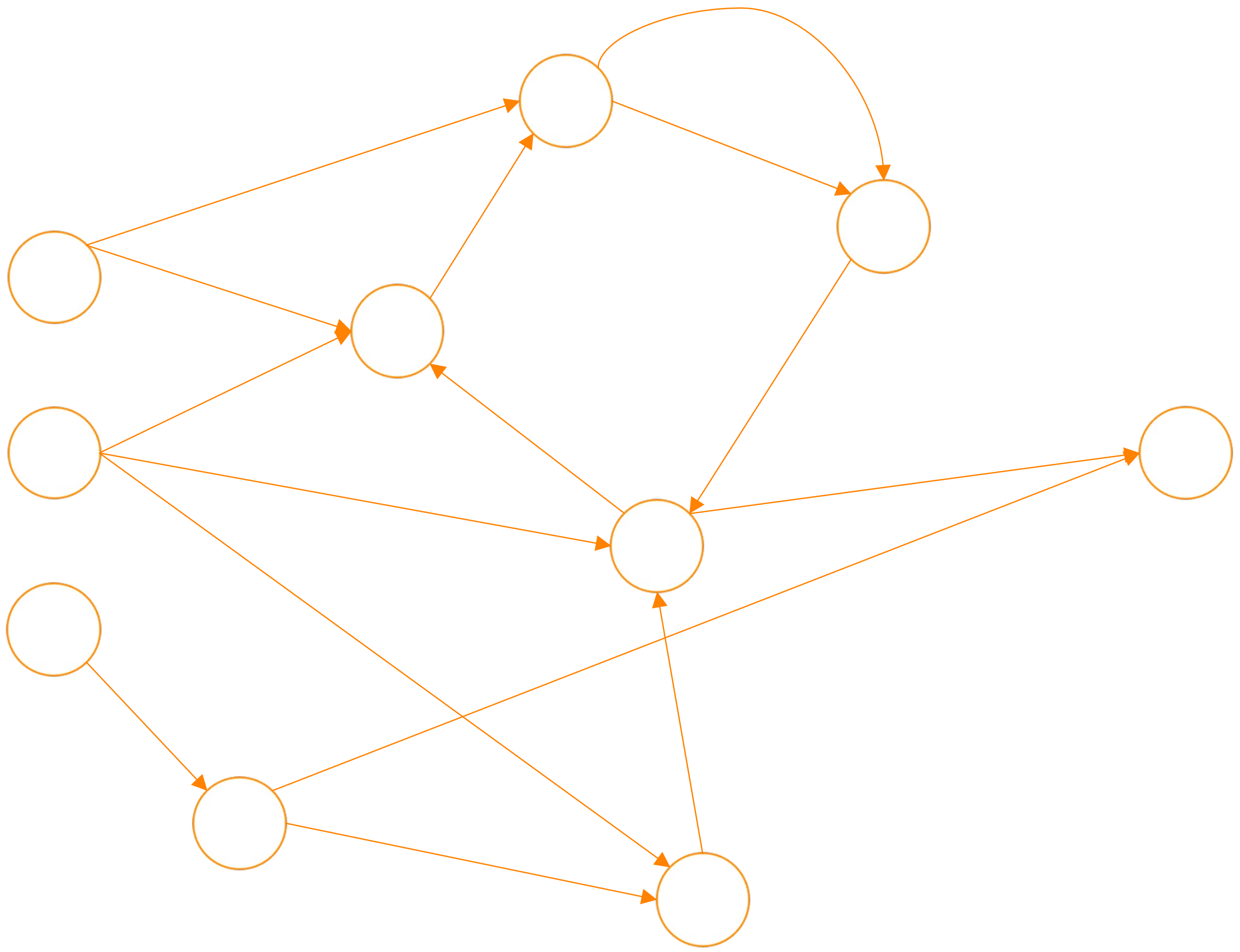


Implements spiking recurrent neural network computation and can be suitable for neuroscience simulation

- Significant promise for future algorithmic development
- Fast computation ***and*** low power
- Still in development

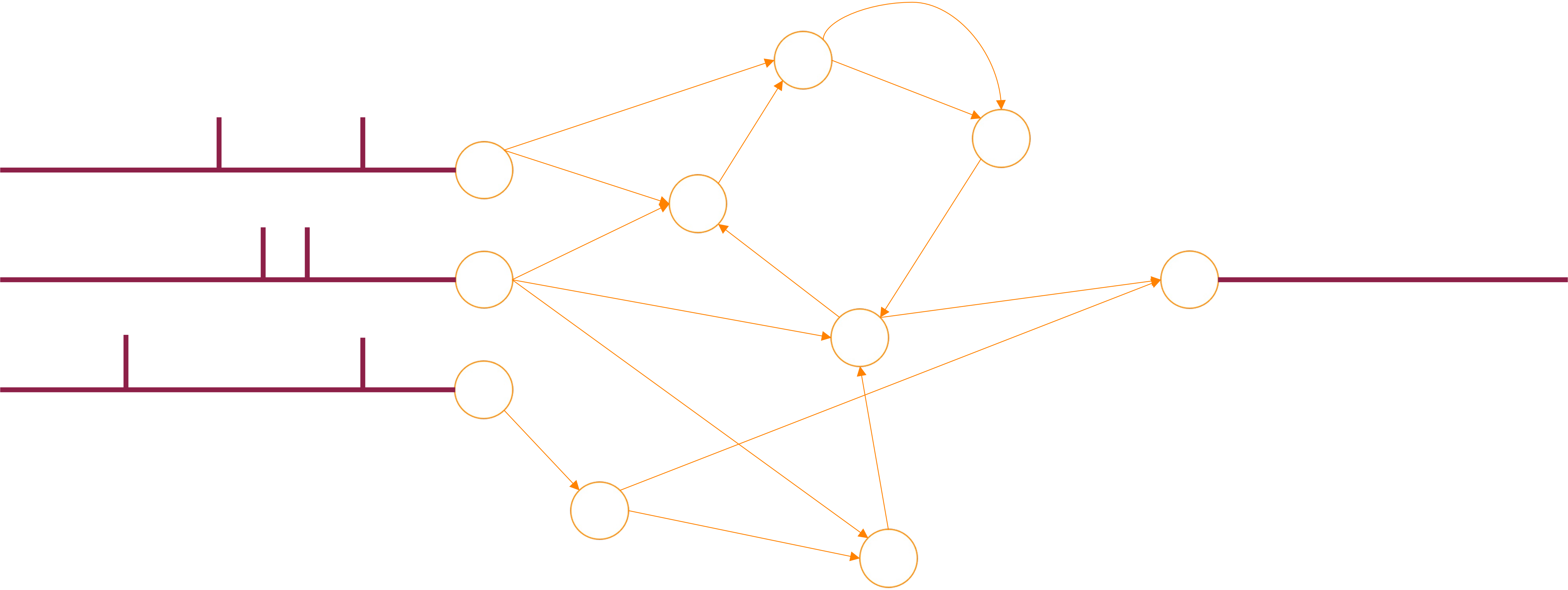
Spiking Neural Networks

- Time component on neurons and synapses
- More complex network structures than feed-forward, but typically not fully connected

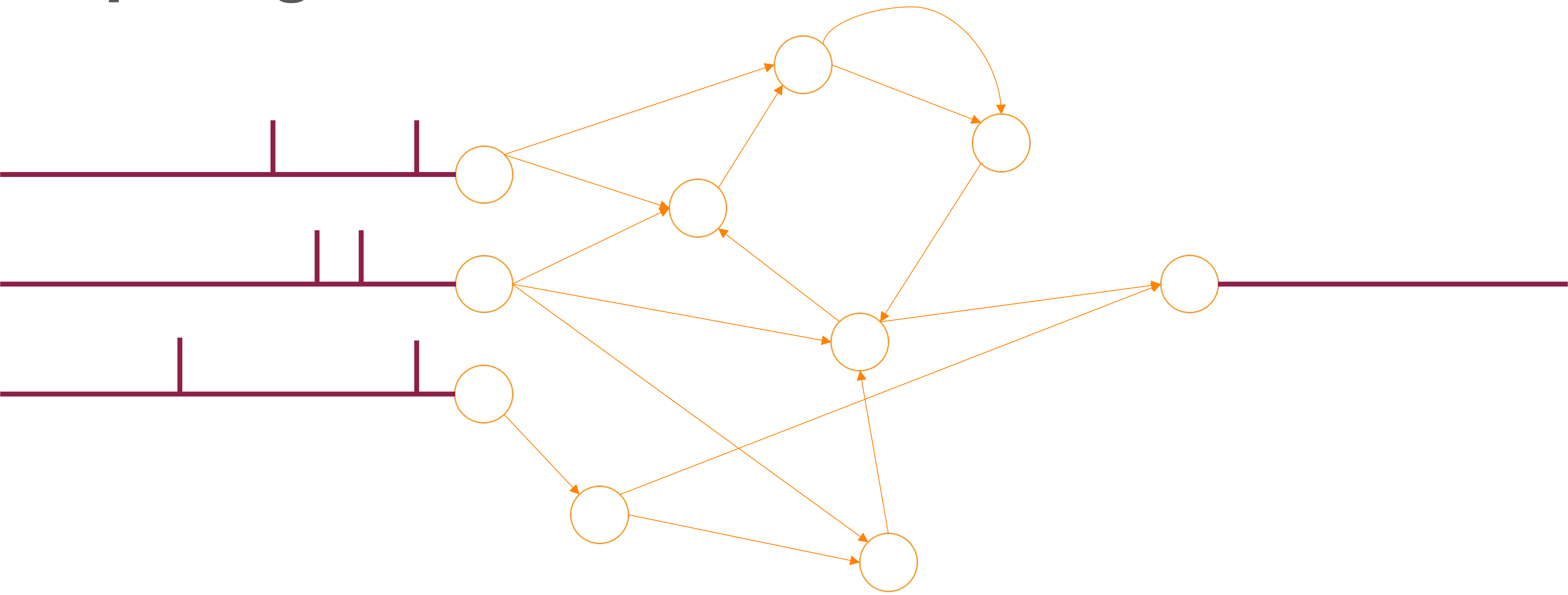


- Temporal input
- Temporal output

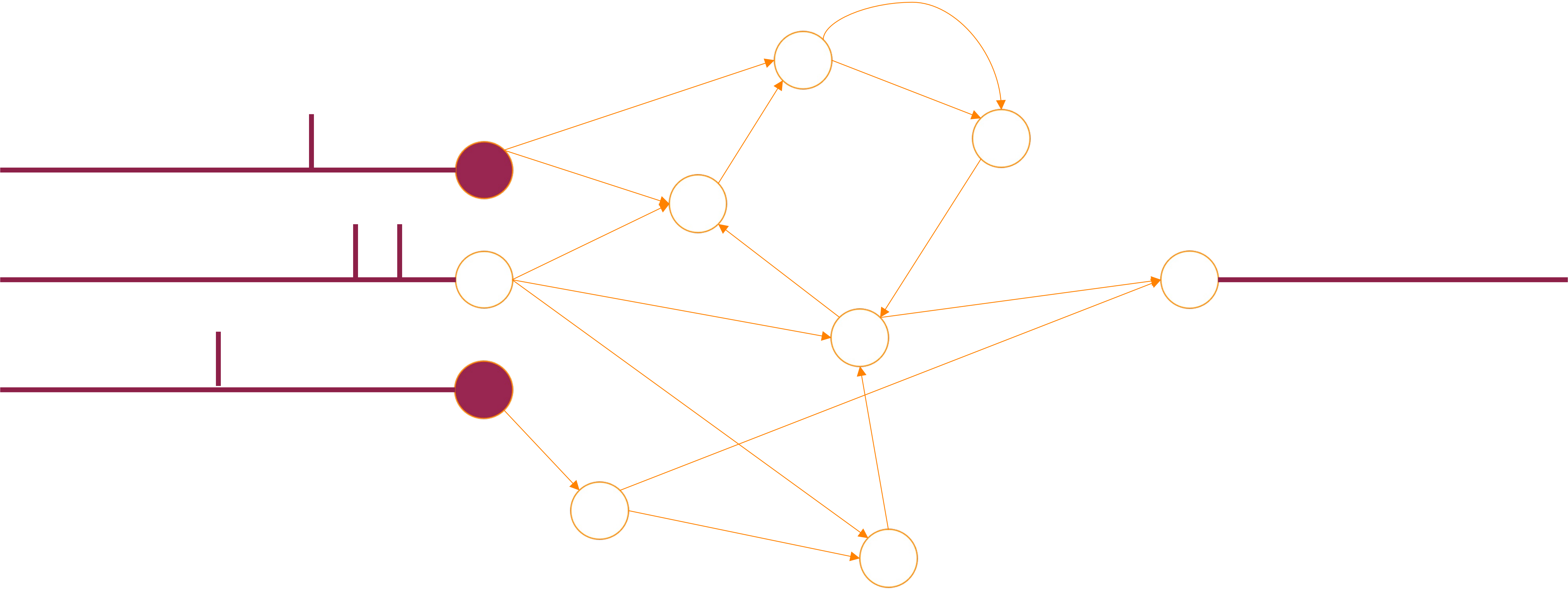
Spiking Neural Networks



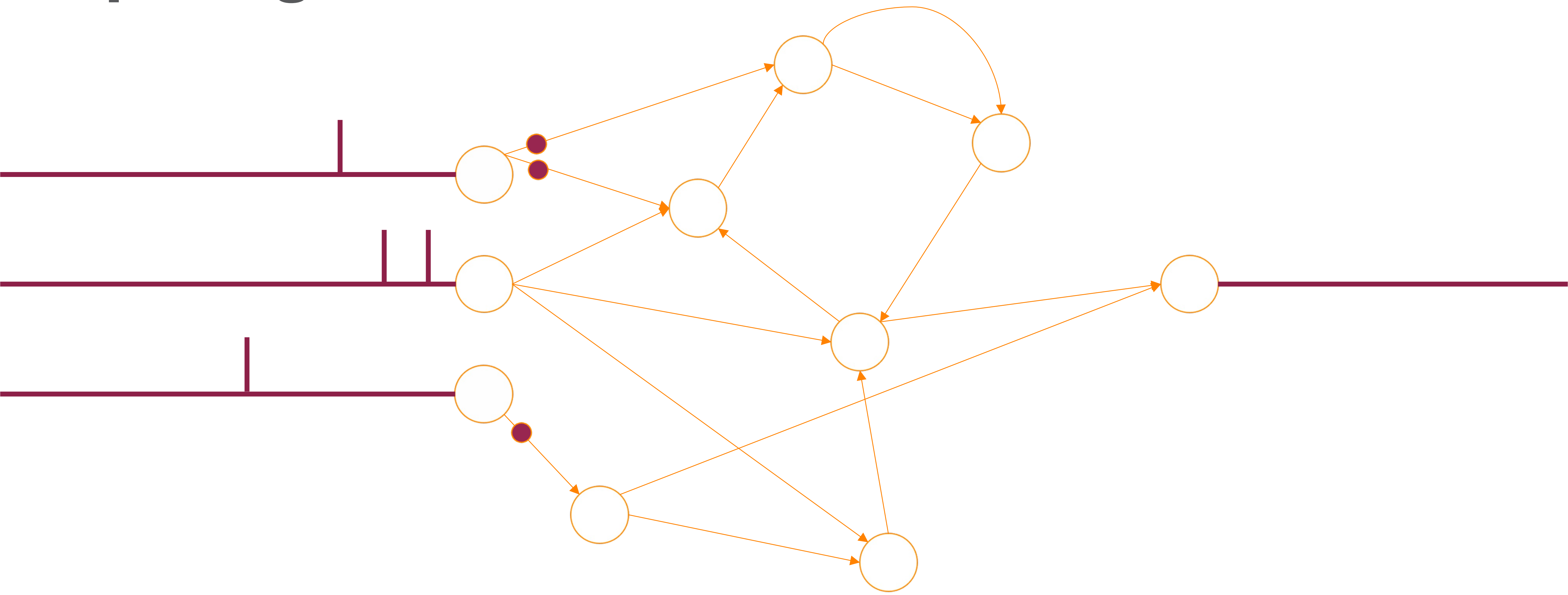
Spiking Neural Networks



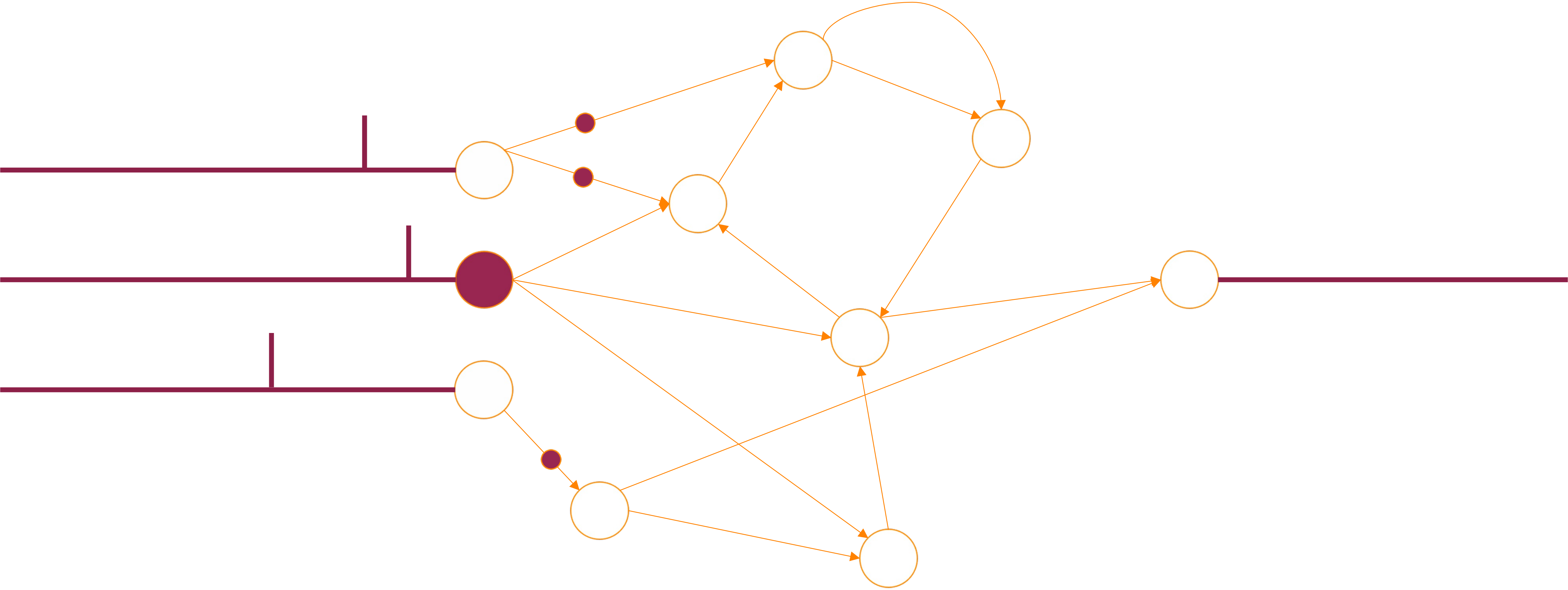
Spiking Neural Networks



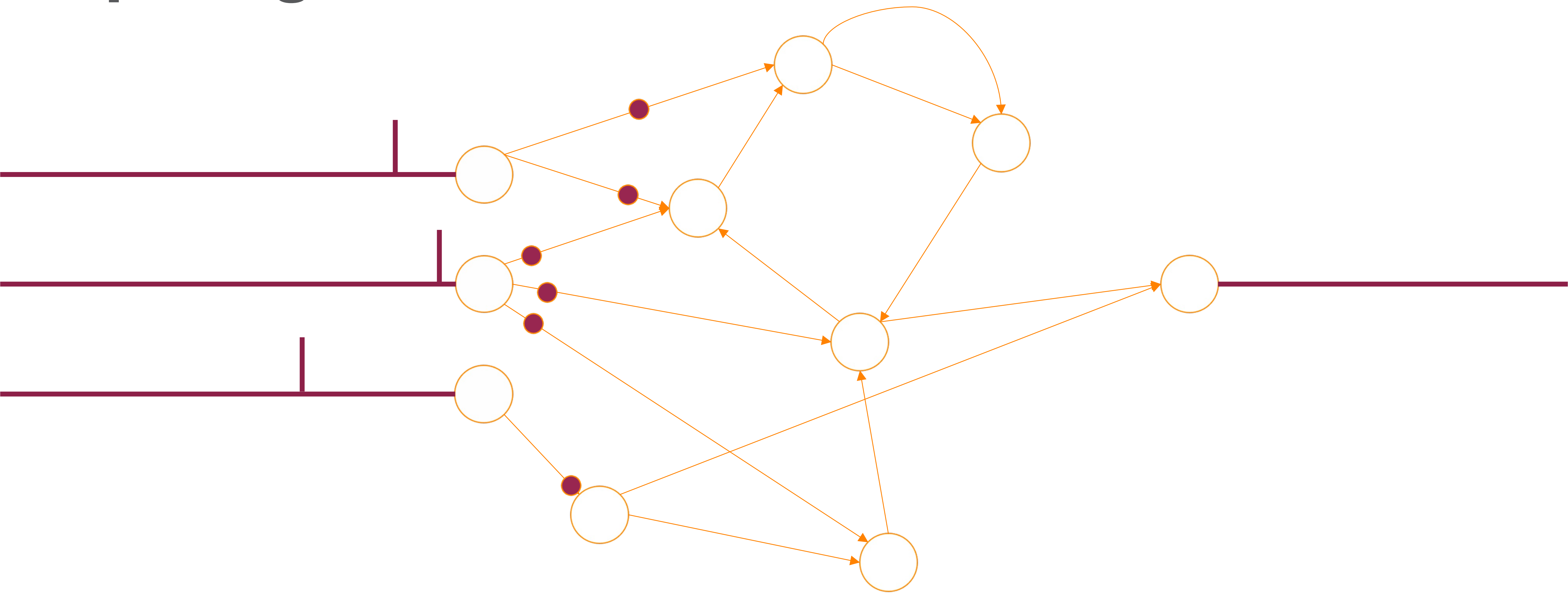
Spiking Neural Networks



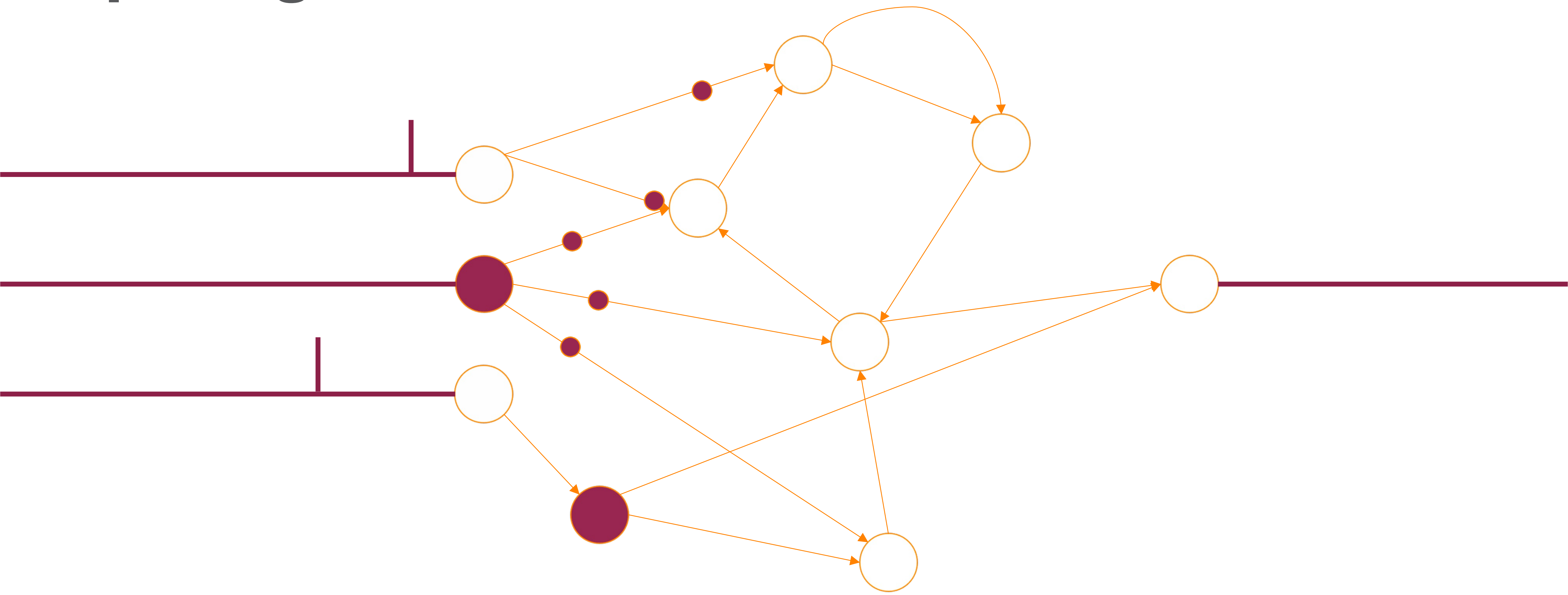
Spiking Neural Networks



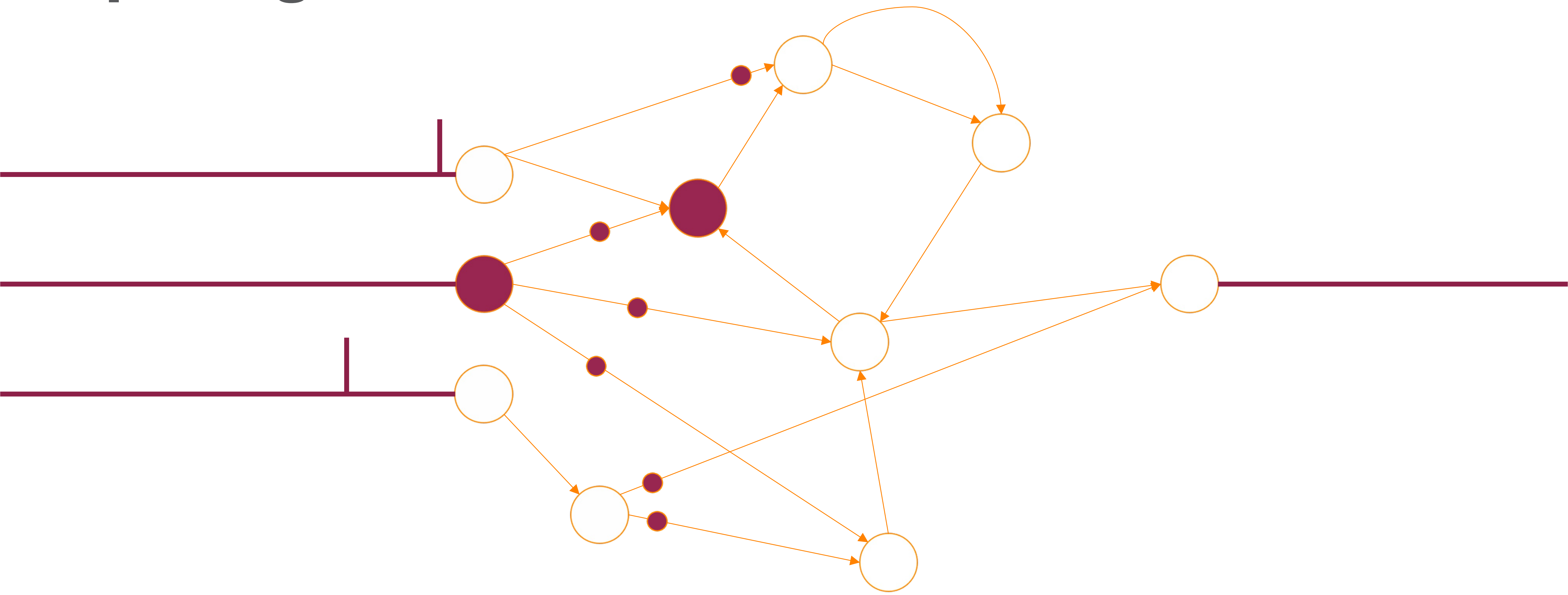
Spiking Neural Networks



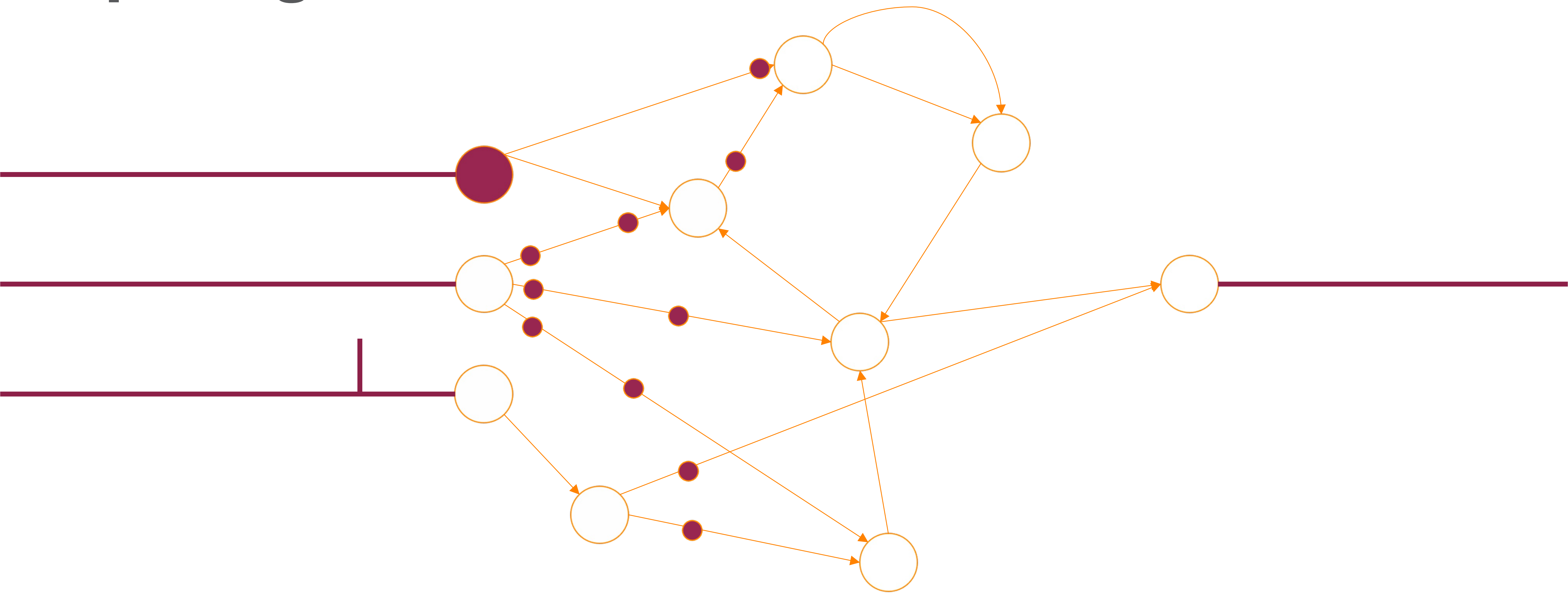
Spiking Neural Networks



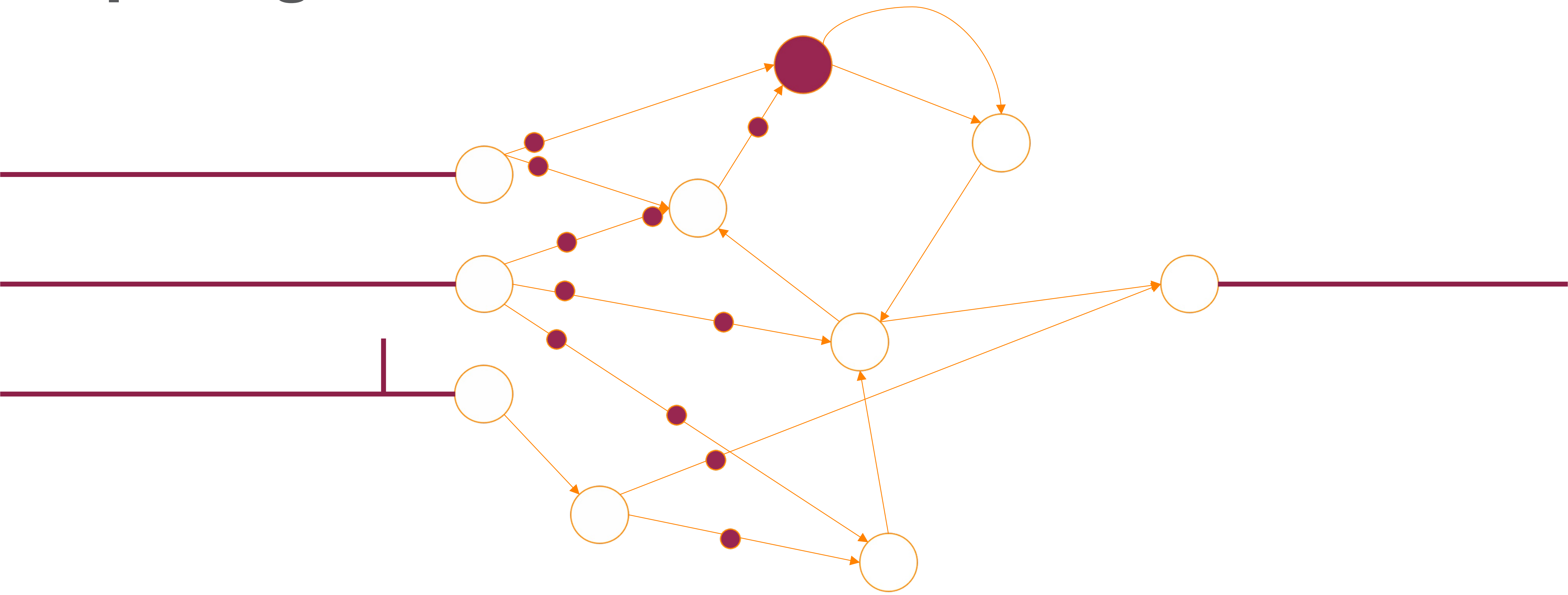
Spiking Neural Networks



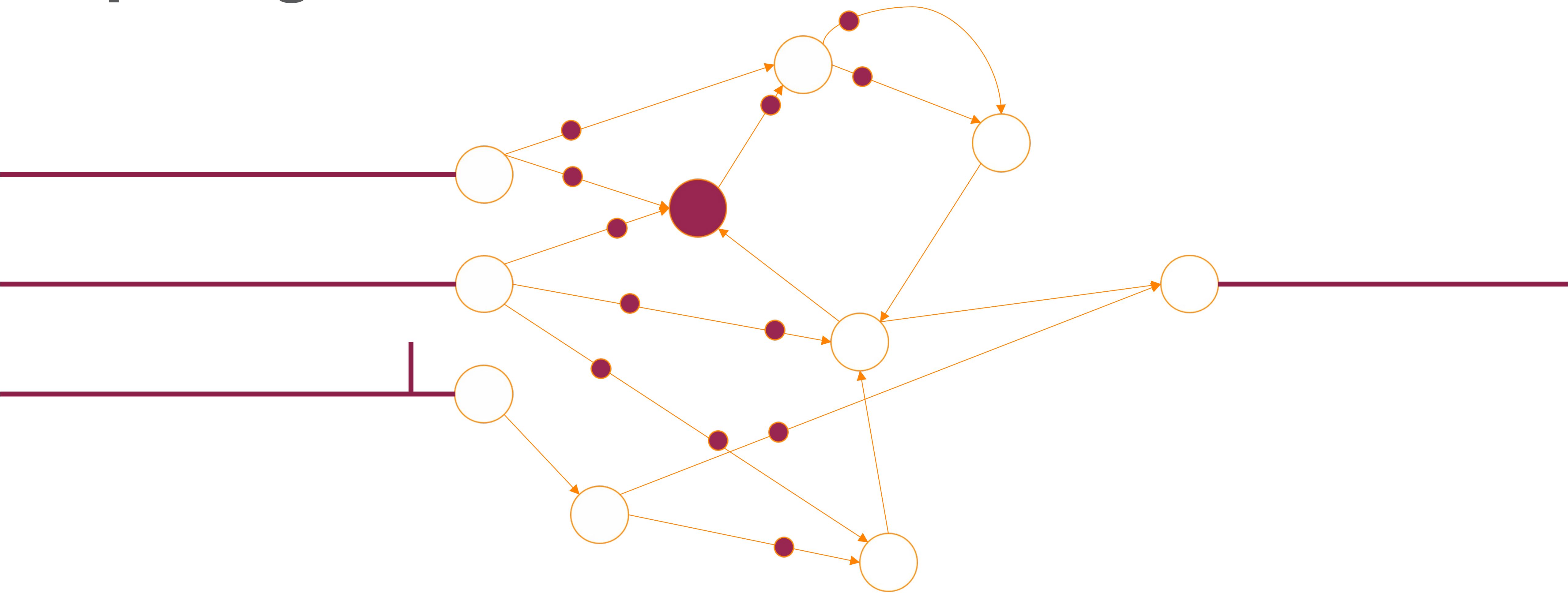
Spiking Neural Networks



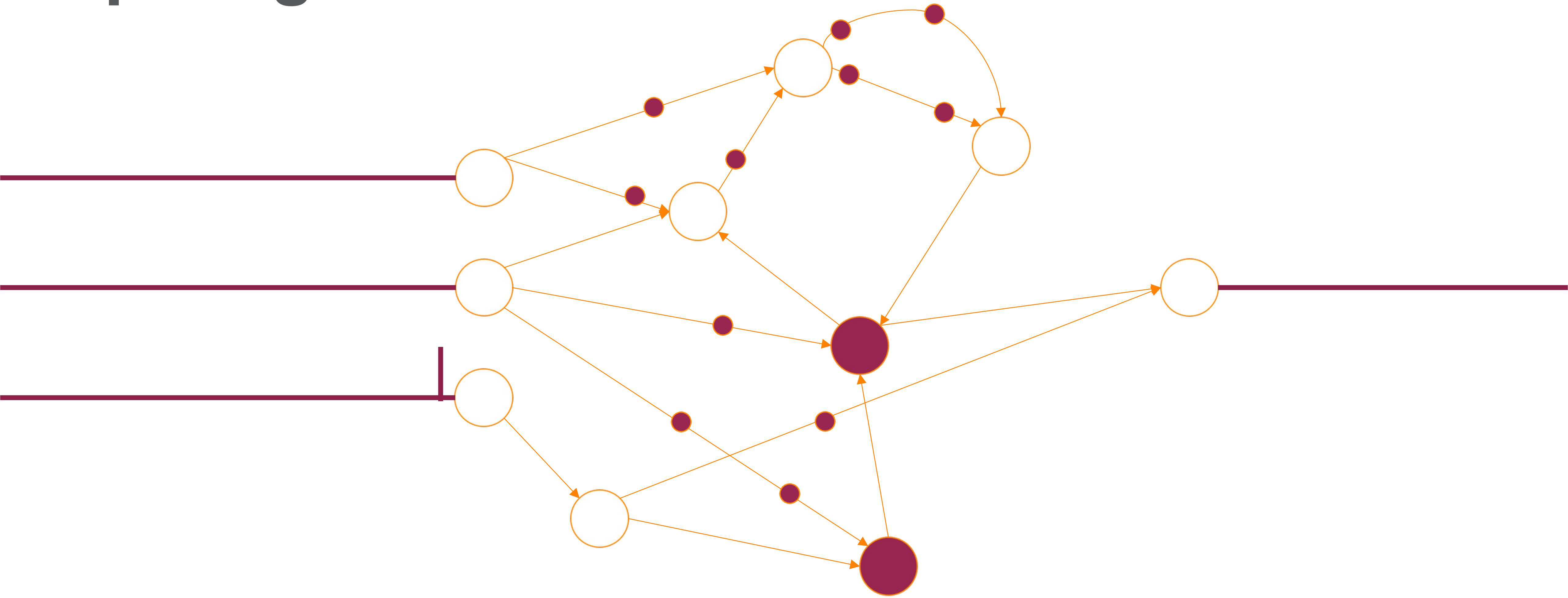
Spiking Neural Networks



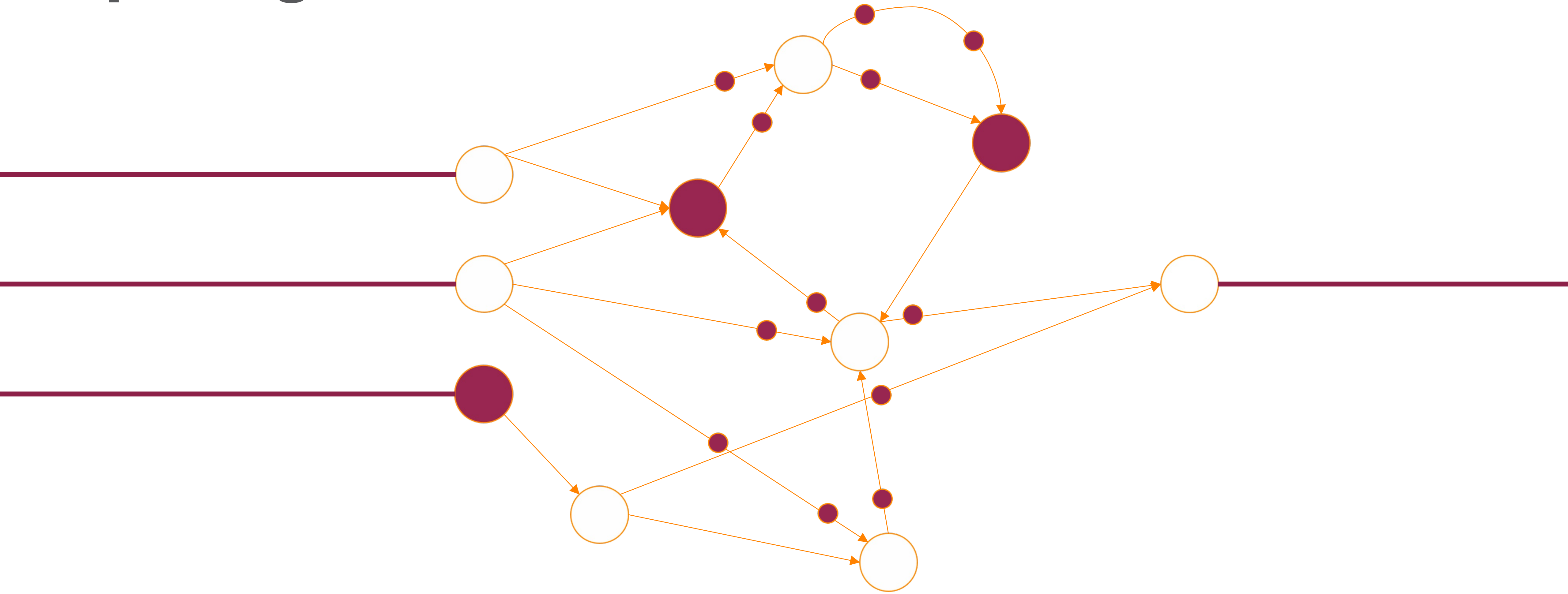
Spiking Neural Networks



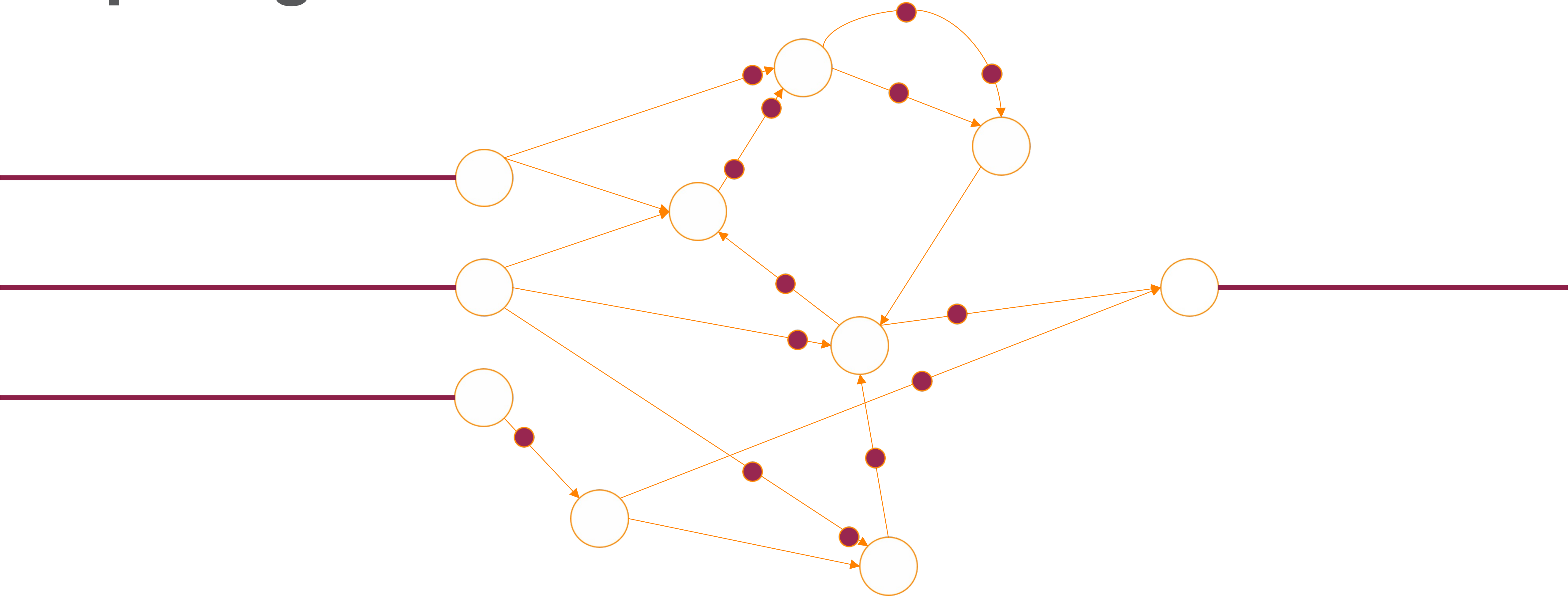
Spiking Neural Networks



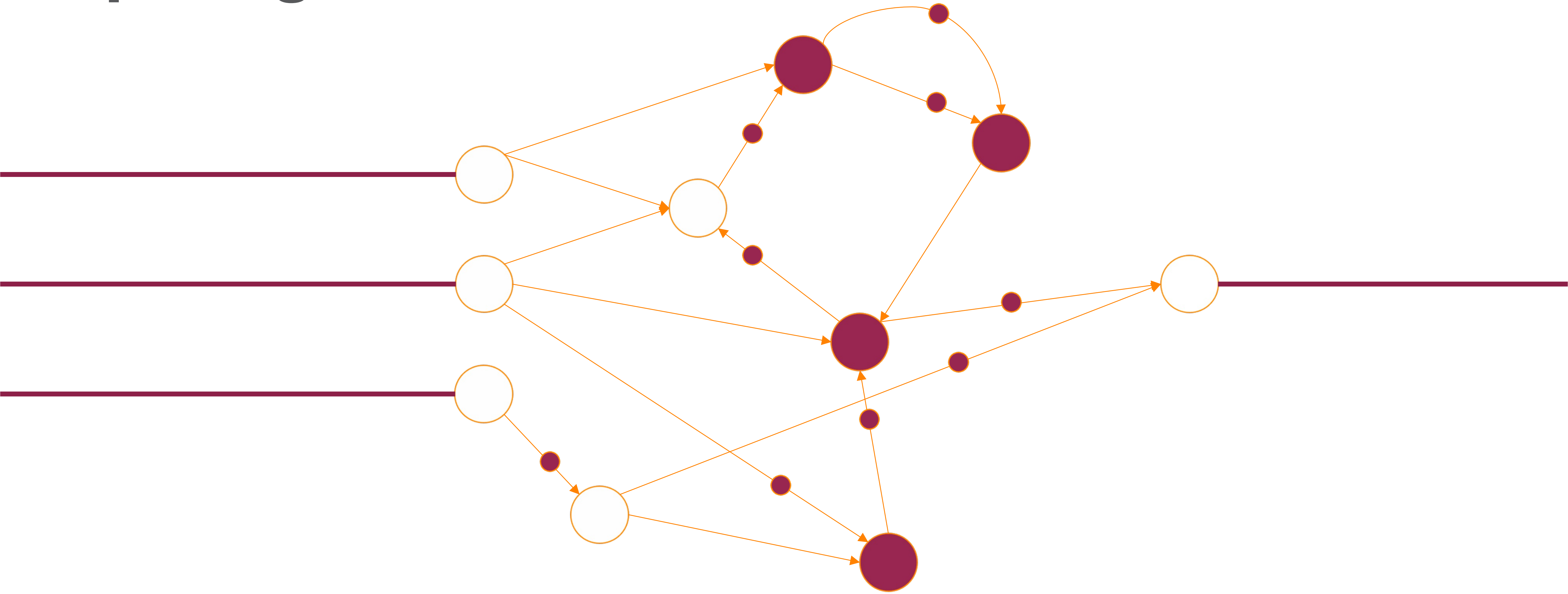
Spiking Neural Networks



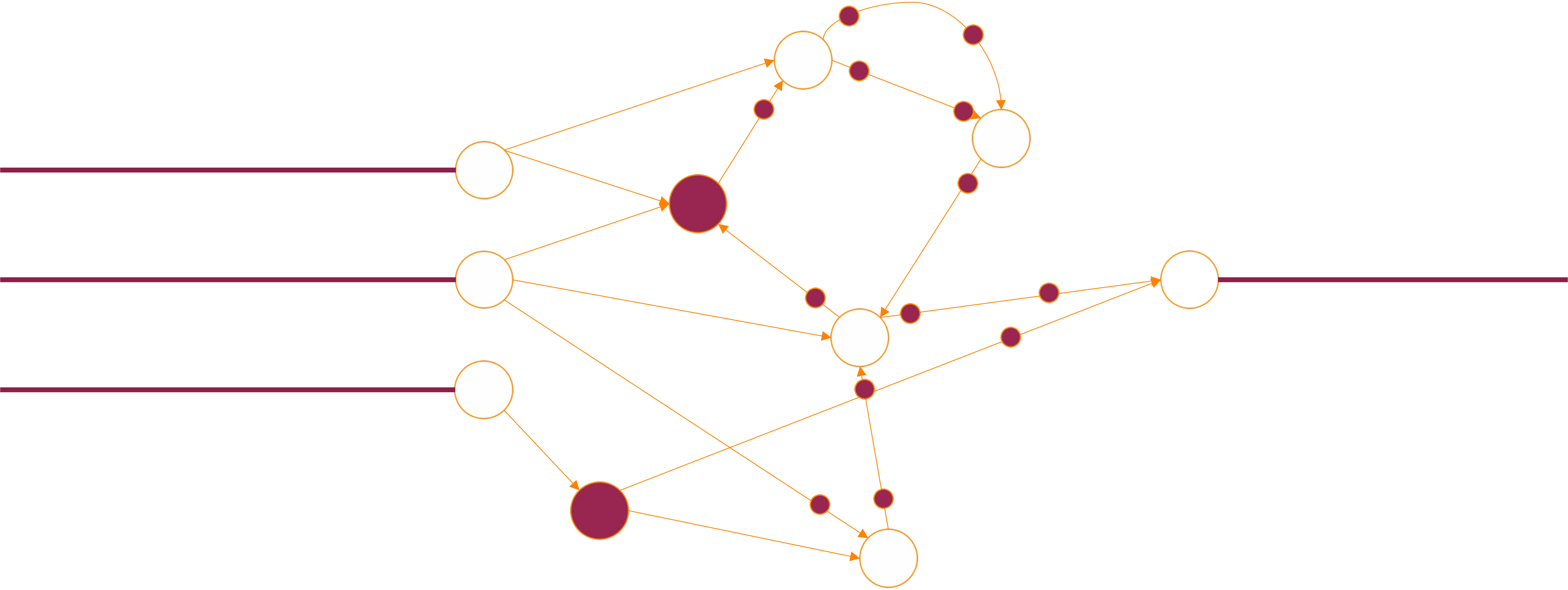
Spiking Neural Networks



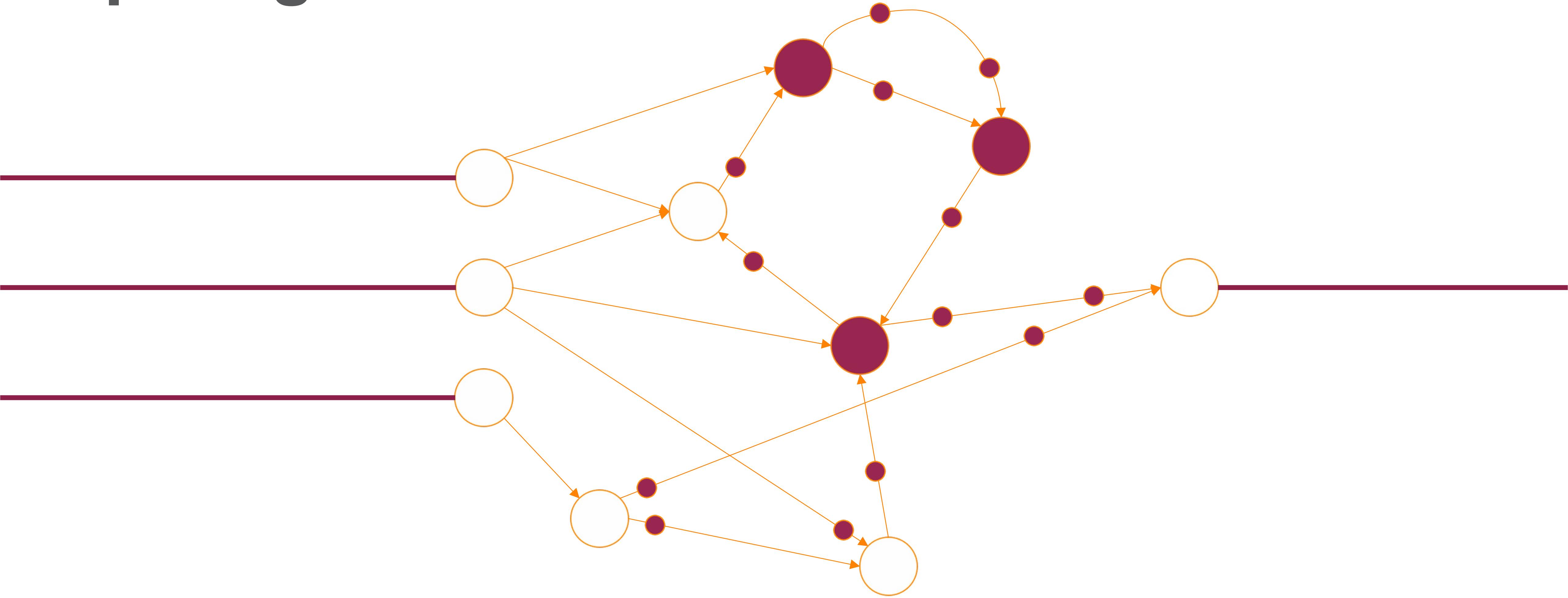
Spiking Neural Networks



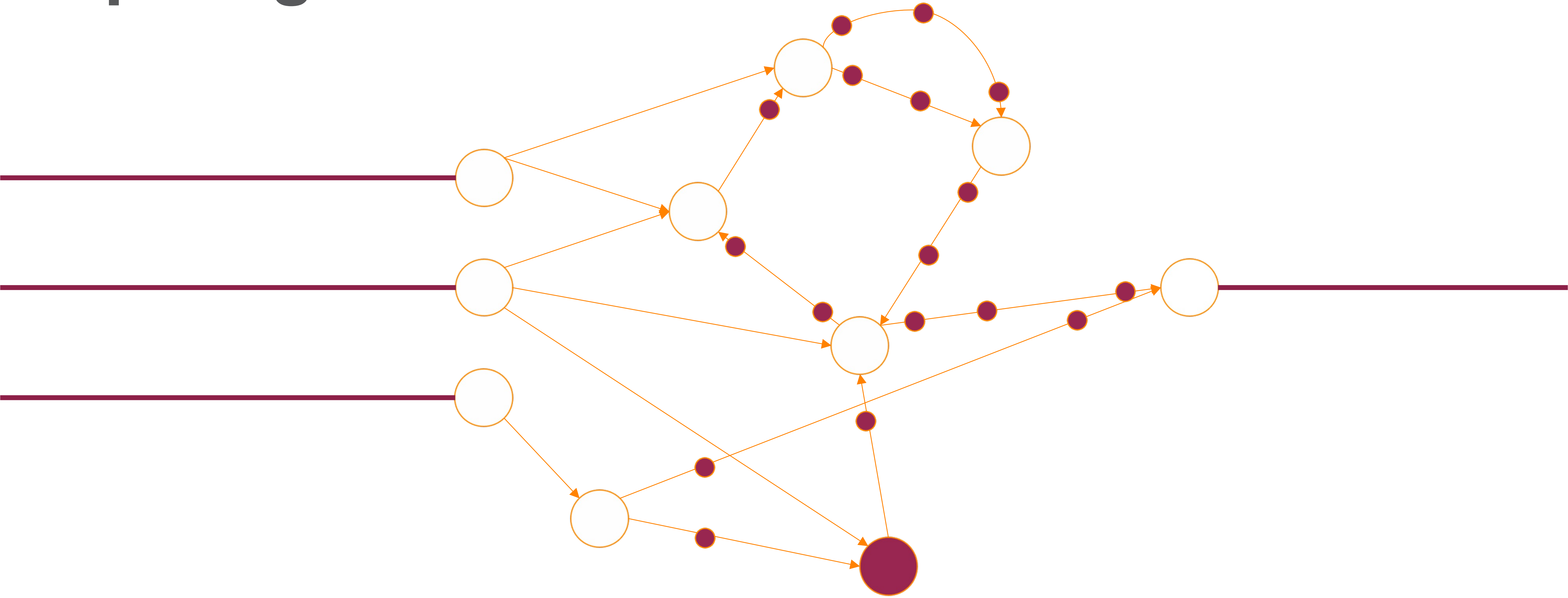
Spiking Neural Networks



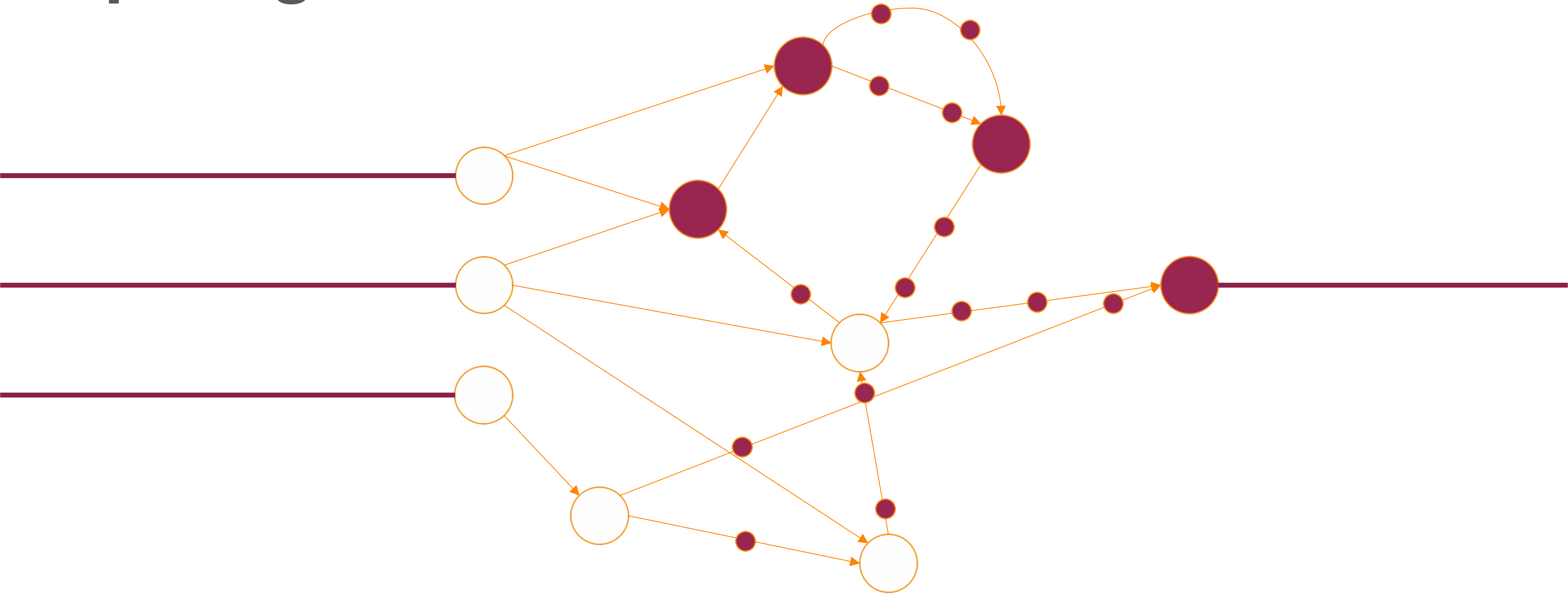
Spiking Neural Networks



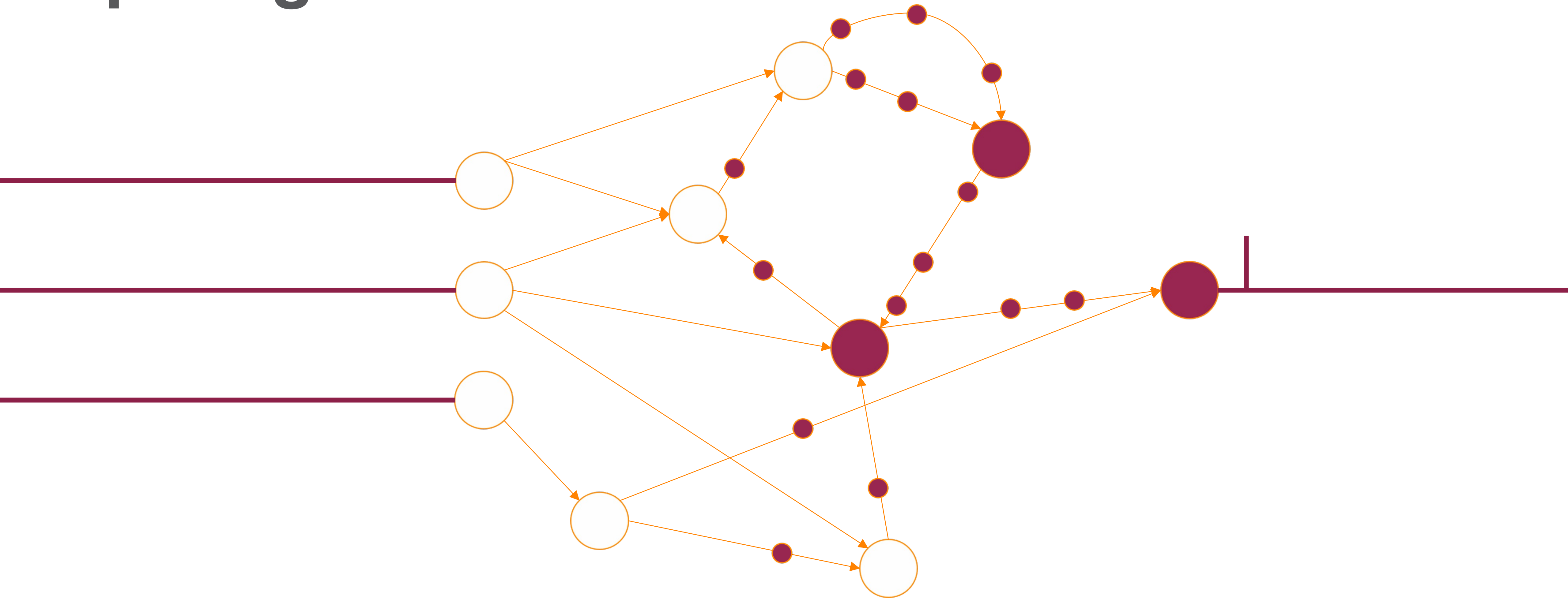
Spiking Neural Networks



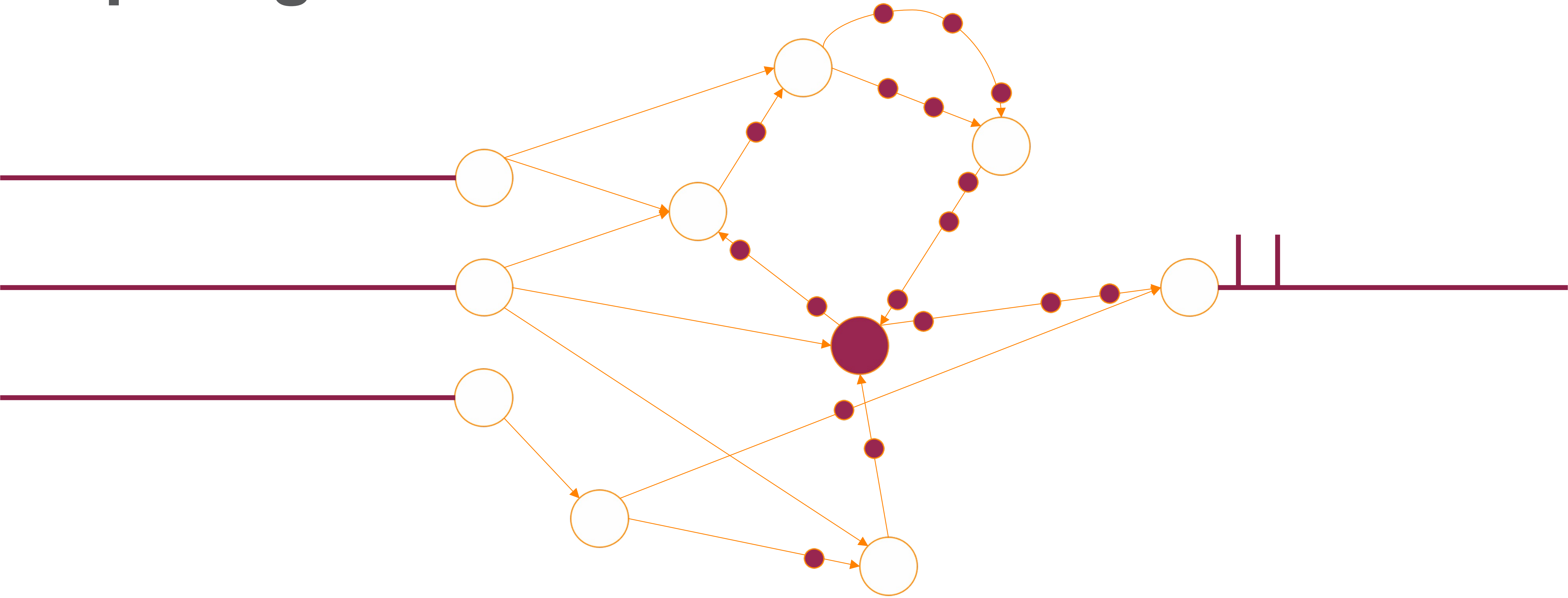
Spiking Neural Networks



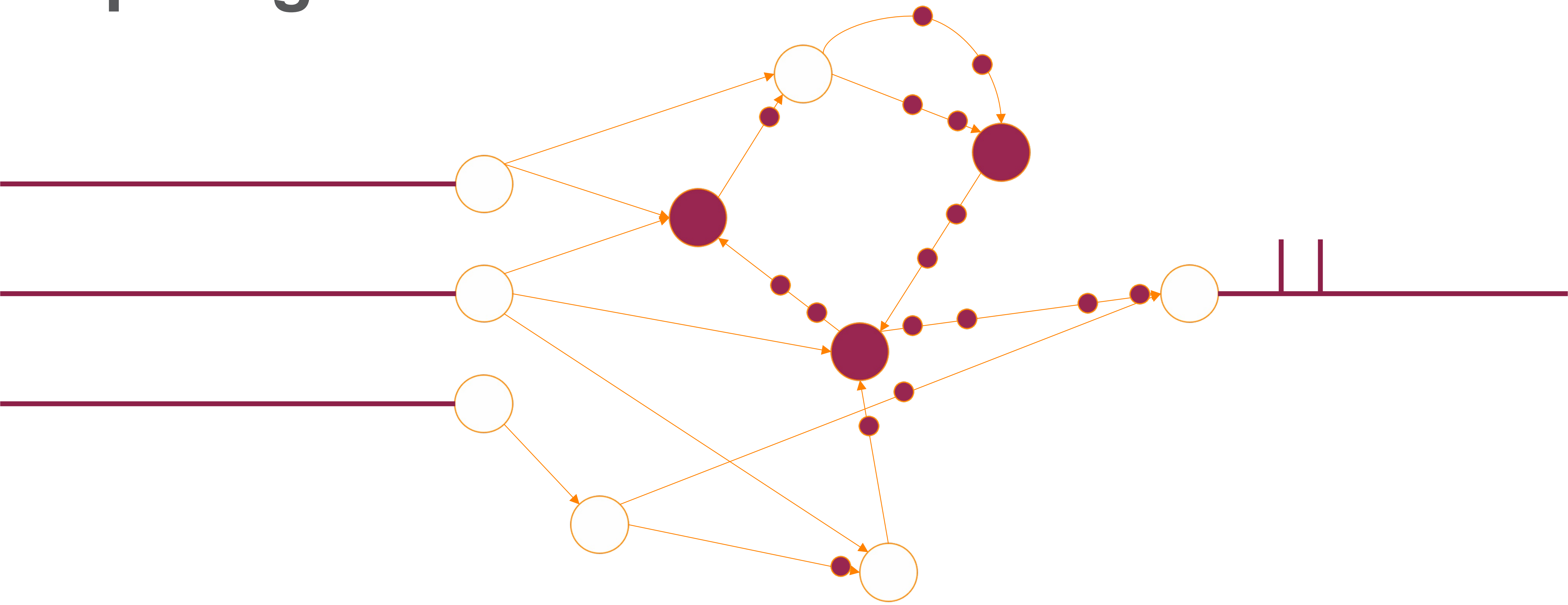
Spiking Neural Networks



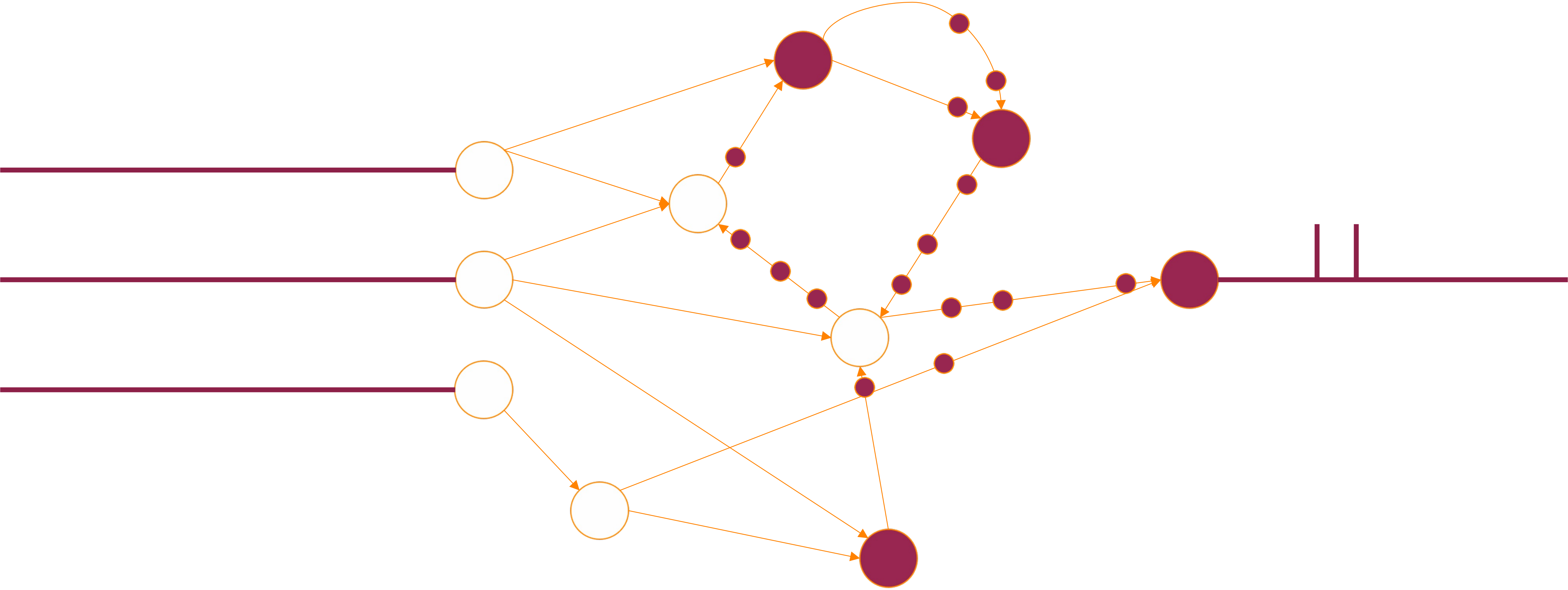
Spiking Neural Networks



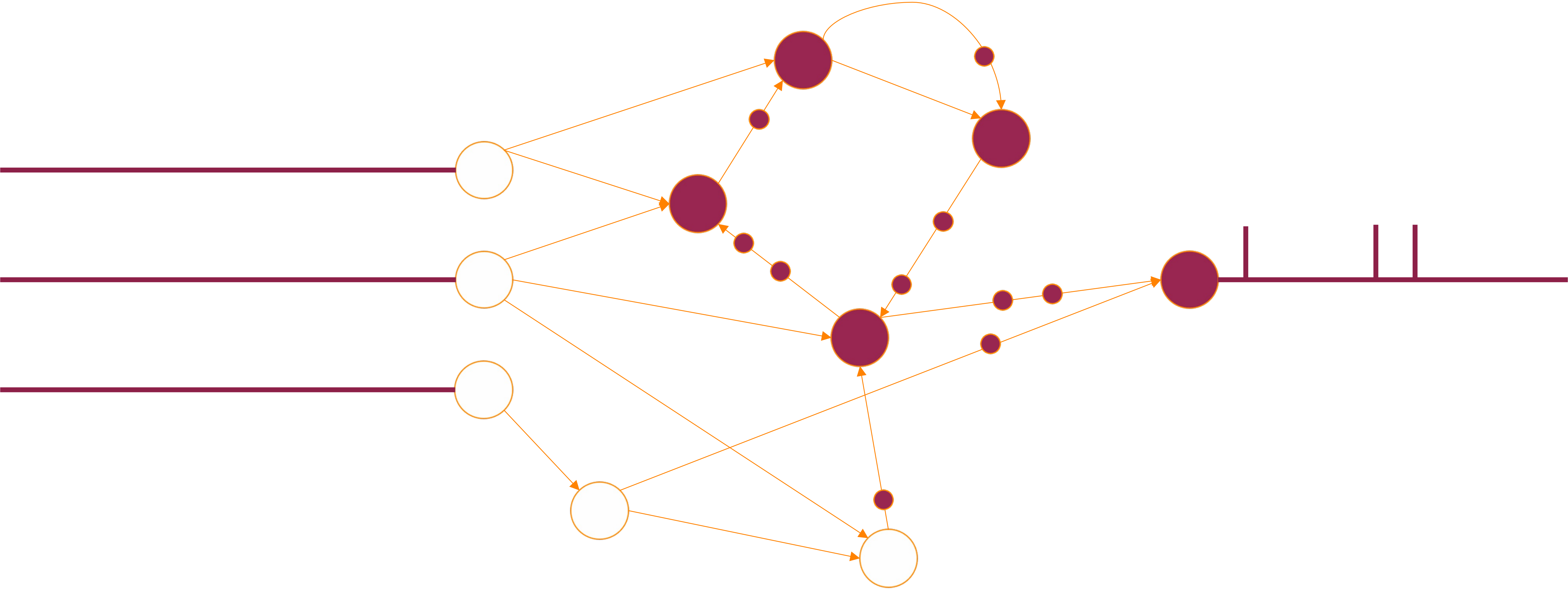
Spiking Neural Networks



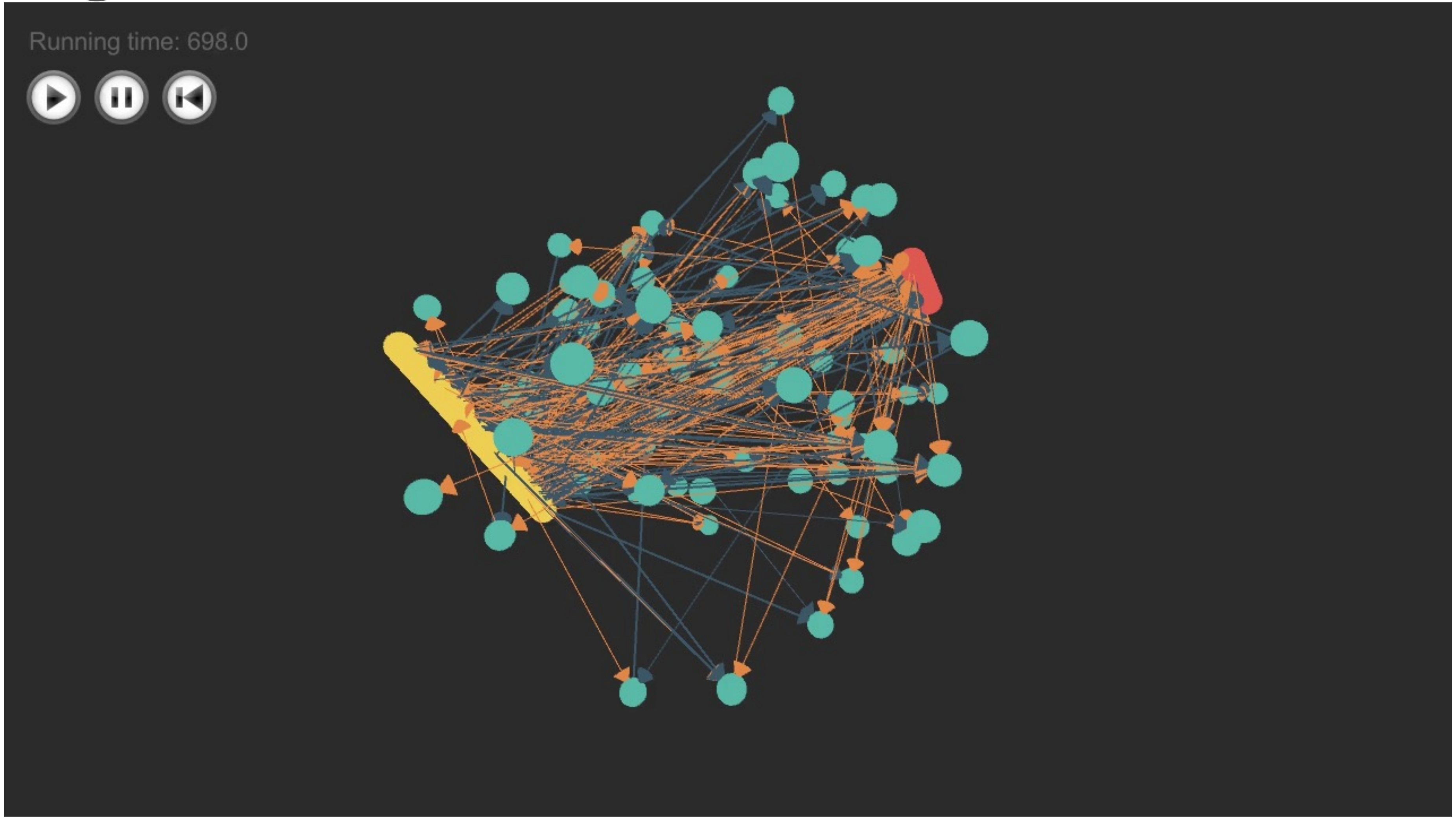
Spiking Neural Networks



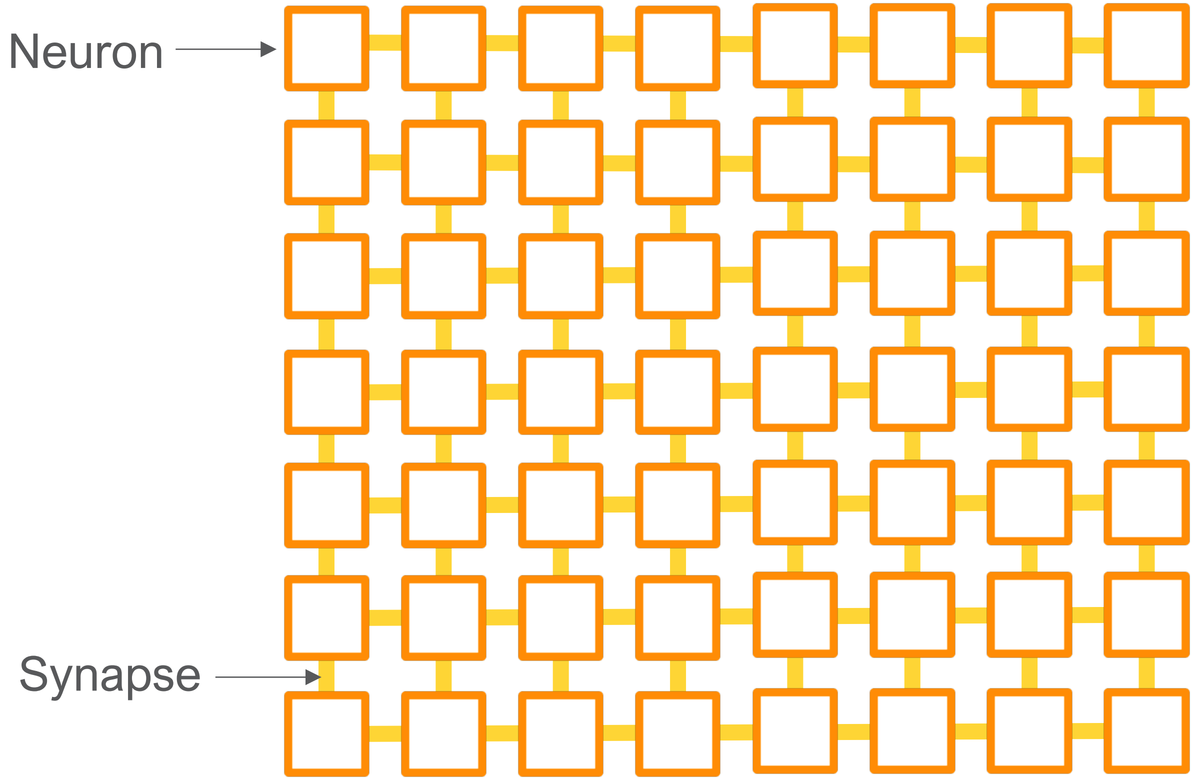
Spiking Neural Networks

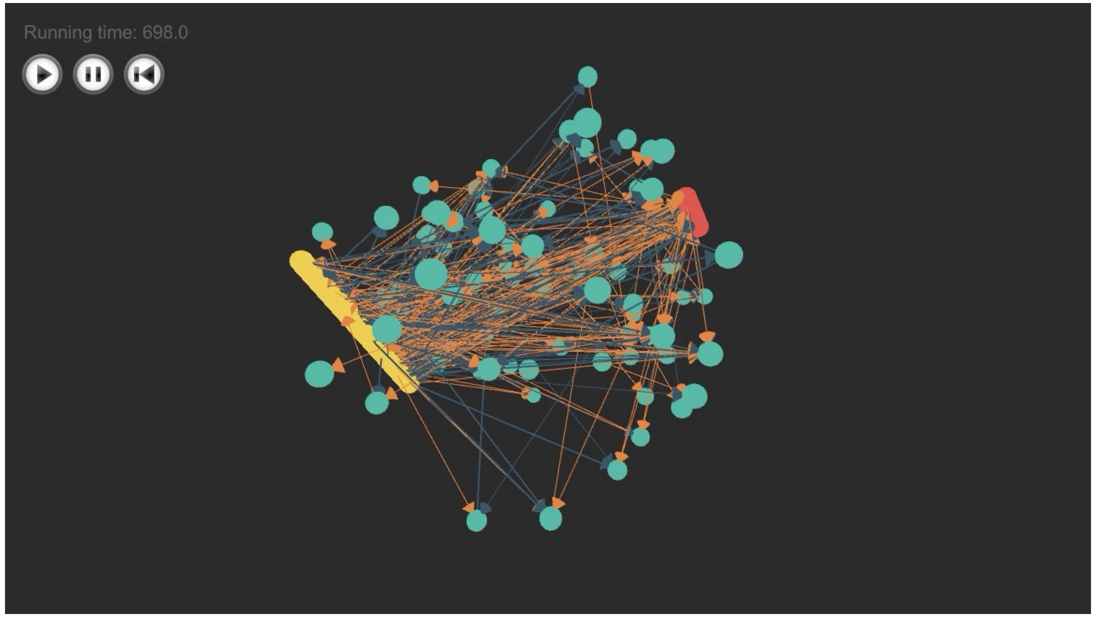


Spiking Neural Networks



What does a neuromorphic computer look like?





Massively Parallel

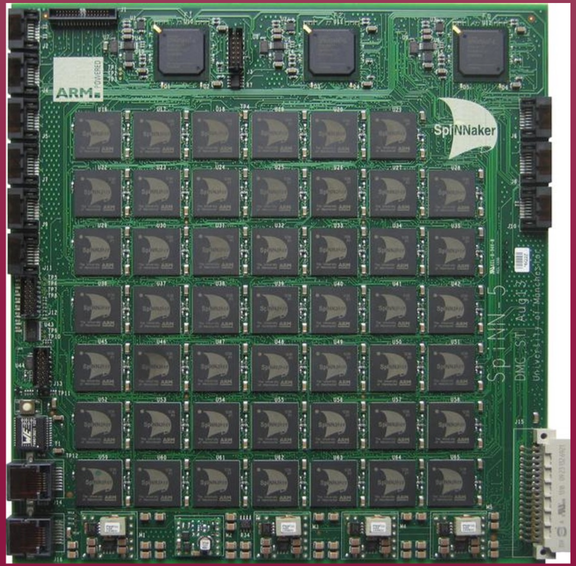
Collocated Processing and Memory

Extremely Low Power

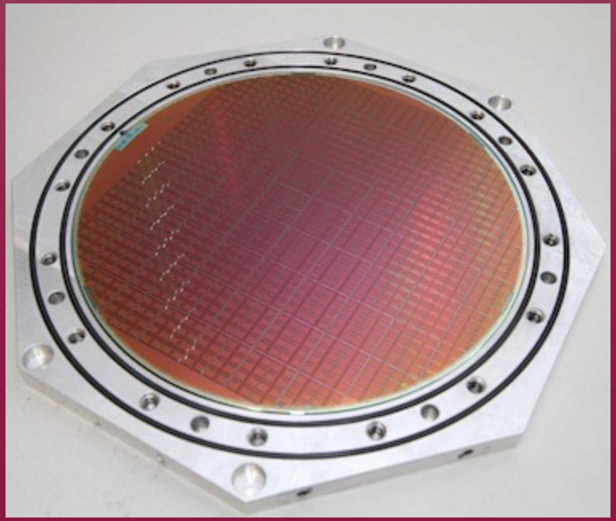
Event Driven

Examples of Neuromorphic Systems

Neuroscience-Driven



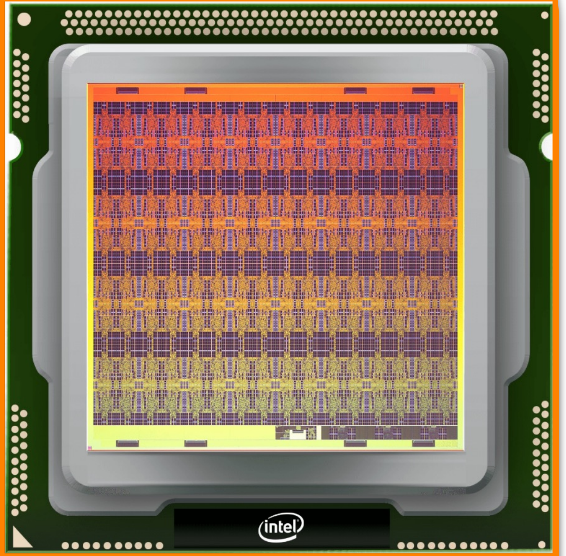
SpiNNaker
University of Manchester



BrainScaleS
Heidelberg University

Computation-Driven

TrueNorth
IBM

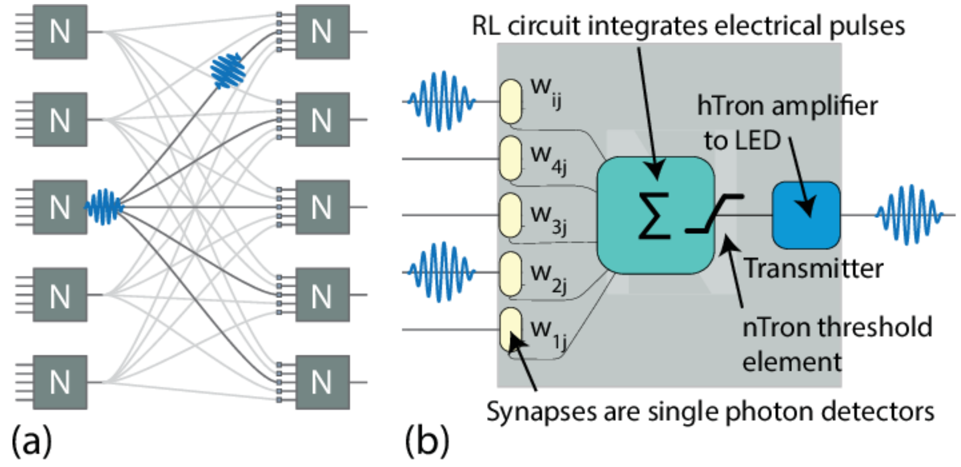


Loihi
Intel

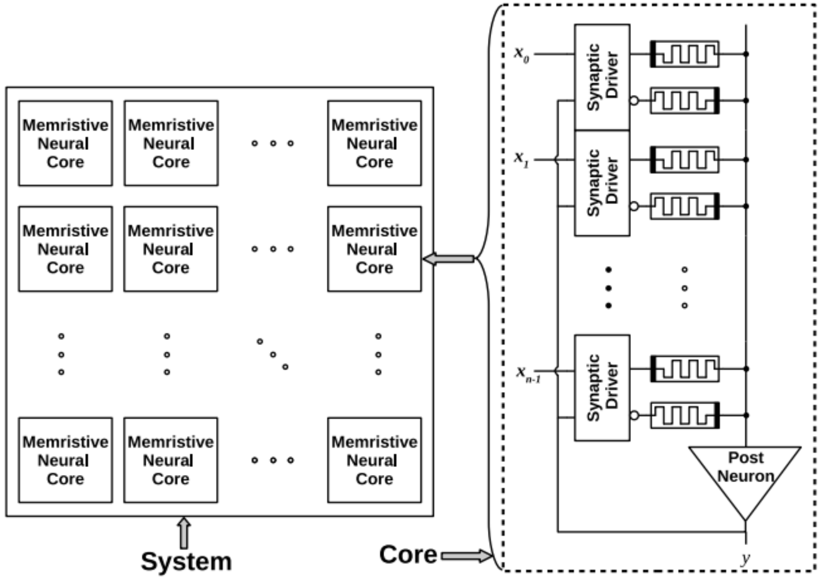
Image Sources:
SpiNNaker: https://www.researchgate.net/figure/A-SpiNNaker-board-with-48-chips-SpiNN-5_fig1_301559712
BrainScaleS: https://www.kip.uni-heidelberg.de/vision/outreach/images/
TrueNorth: https://www.top500.org/news/ibm-finds-killer-app-for-truenorth-neuromorphic-chip/
Loihi: https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html

Neuromorphic Hardware Research

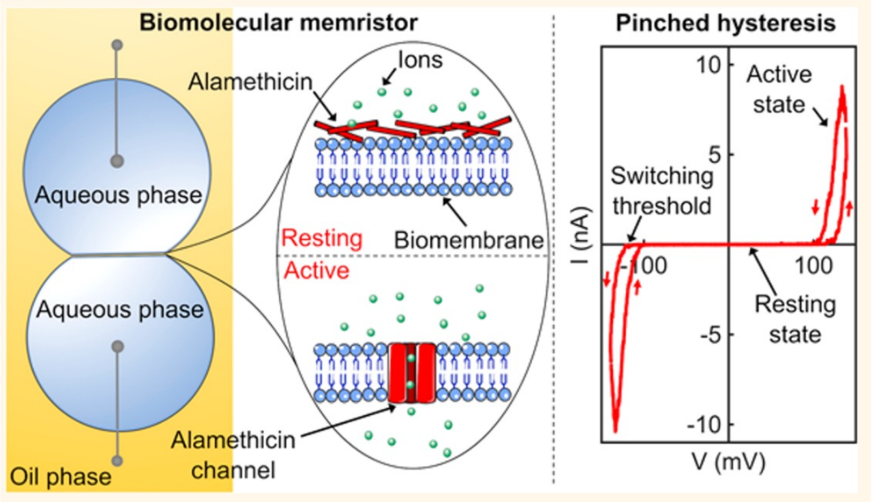
Neuromorphic device research includes
metal-oxide memristors,
superconducting optoelectronics, and biomimetic devices



Buckley, Sonia, et al. "Design of superconducting optoelectronic networks for neuromorphic computing." In *2018 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1-7. IEEE, 2018.



G. Chakma, et al, "Memristive Mixed-Signal Neuromorphic Systems: Energy-Efficient Learning at the Circuit-Level," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 125-136, March 2018.



Najem, Joseph S., et al. "Memristive ion channel-doped biomembranes as synaptic mimics." *ACS nano* 12, no. 5 (2018): 4702-4711.

Neuromorphic Computing “Stack”

Applications

Algorithms

System Software and Communications

System Architecture/Organization

Microarchitecture

Devices

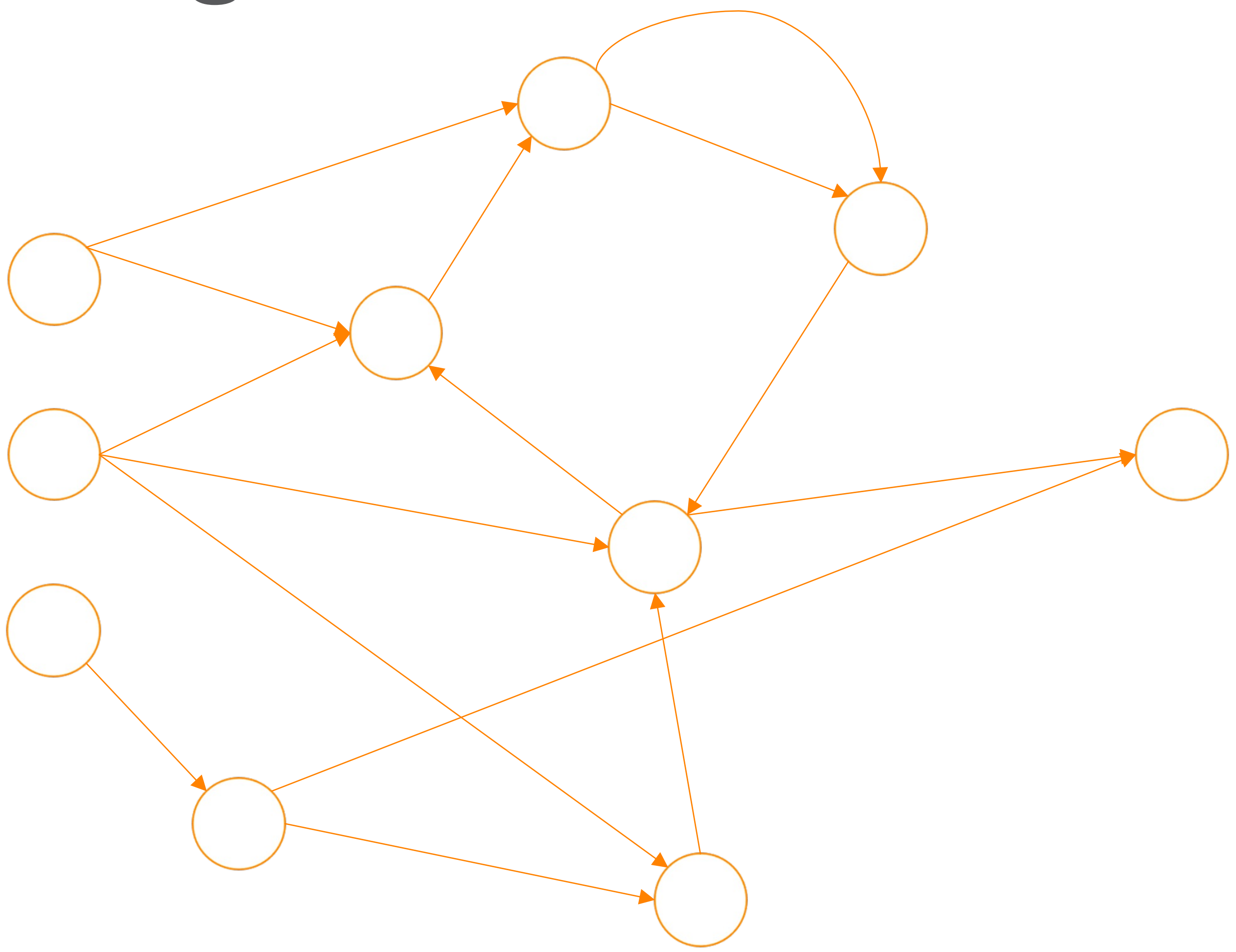
Materials

My Research

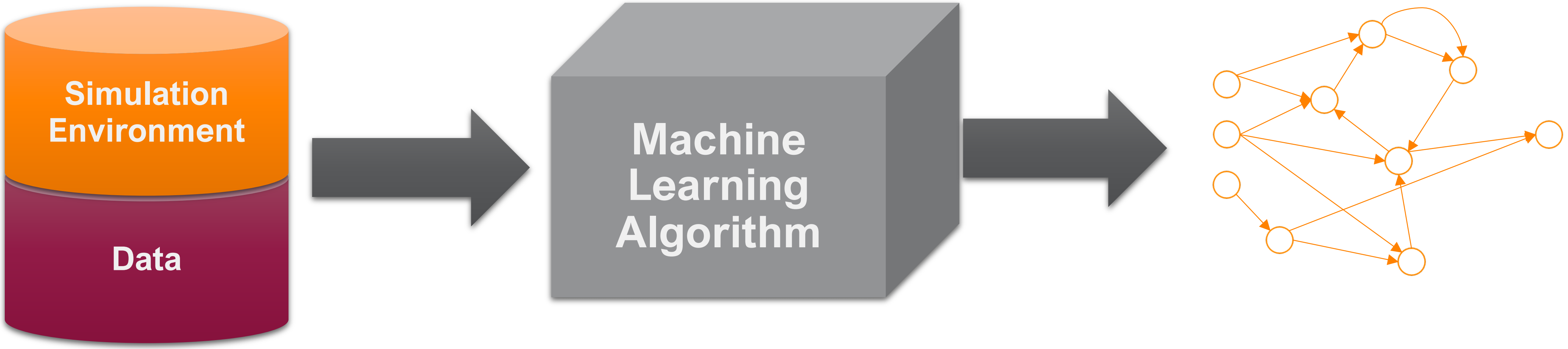
Influences

How do you program a neuromorphic computer?

Spiking Neural Networks!

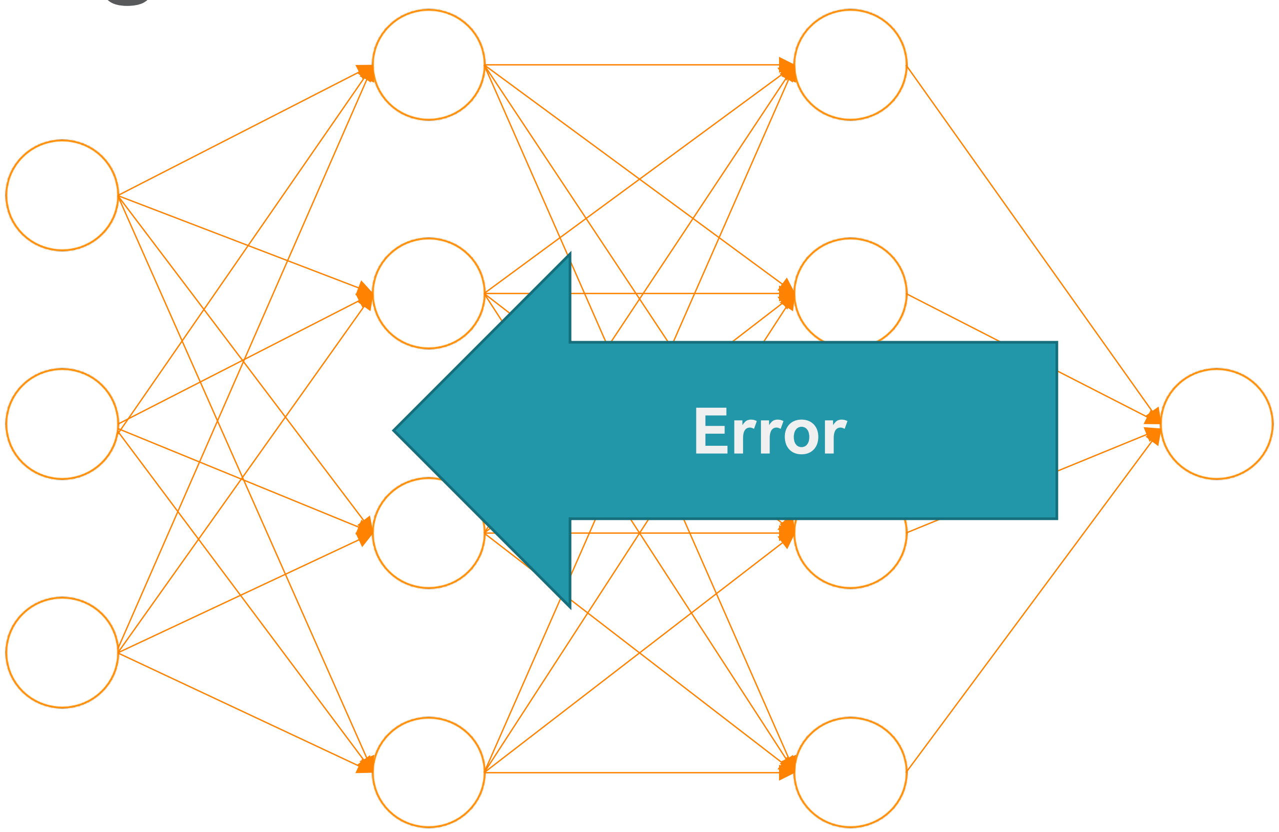


“Programming” via Spiking Neural Networks



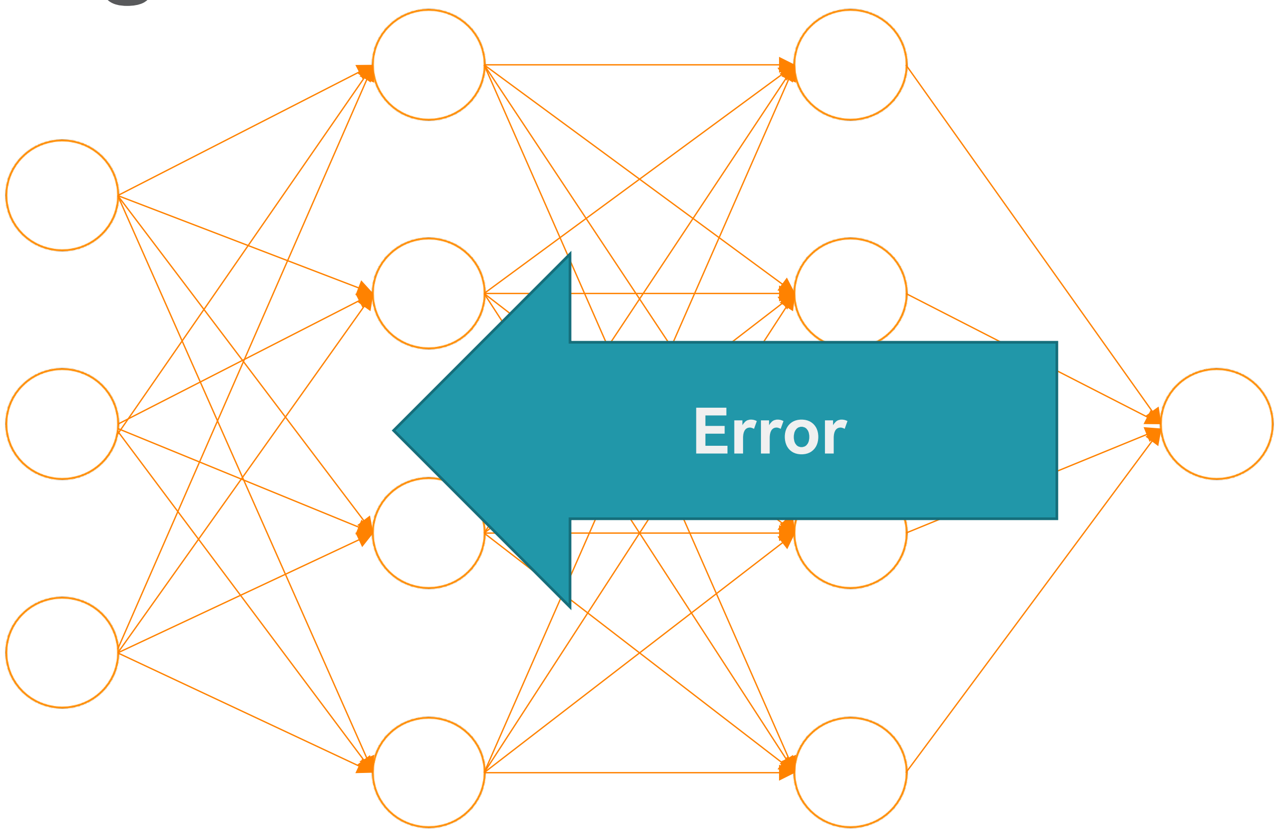
Algorithms: Back-Propagation-Like Approaches

- Dense connectivity
- Algorithm adaptations for:
  - **Non-differentiability of spiking neurons**
  - Low precision weights
  - Non-standard approach to delays



Algorithms: Back-Propagation-Like Approaches

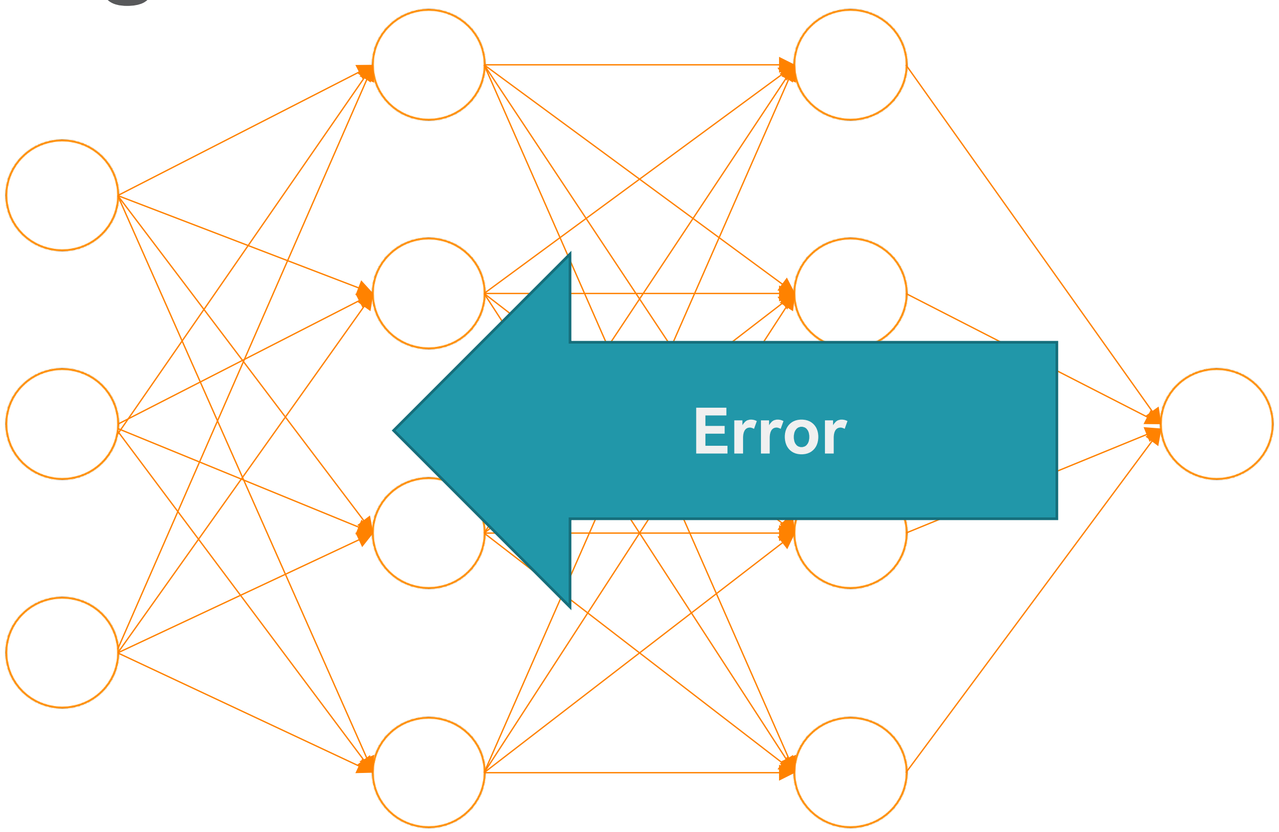
- Dense connectivity
- Algorithm adaptations for:
  - **Non-differentiability of spiking neurons**
  - Low precision weights
  - Non-standard approach to delays



Key Advantage:
Decades of knowledge about traditional ANNs

Algorithms: Back-Propagation-Like Approaches

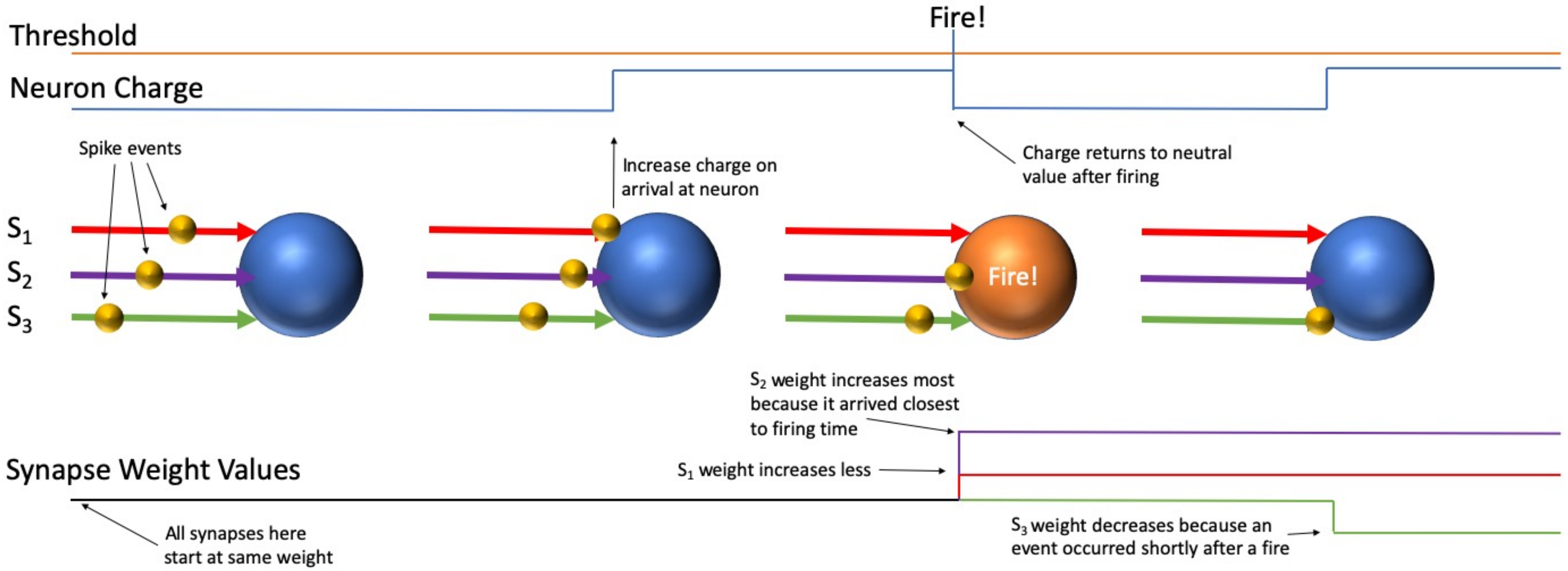
- Dense connectivity
- Algorithm adaptations for:
  - **Non-differentiability of spiking neurons**
  - Low precision weights
  - Non-standard approach to delays



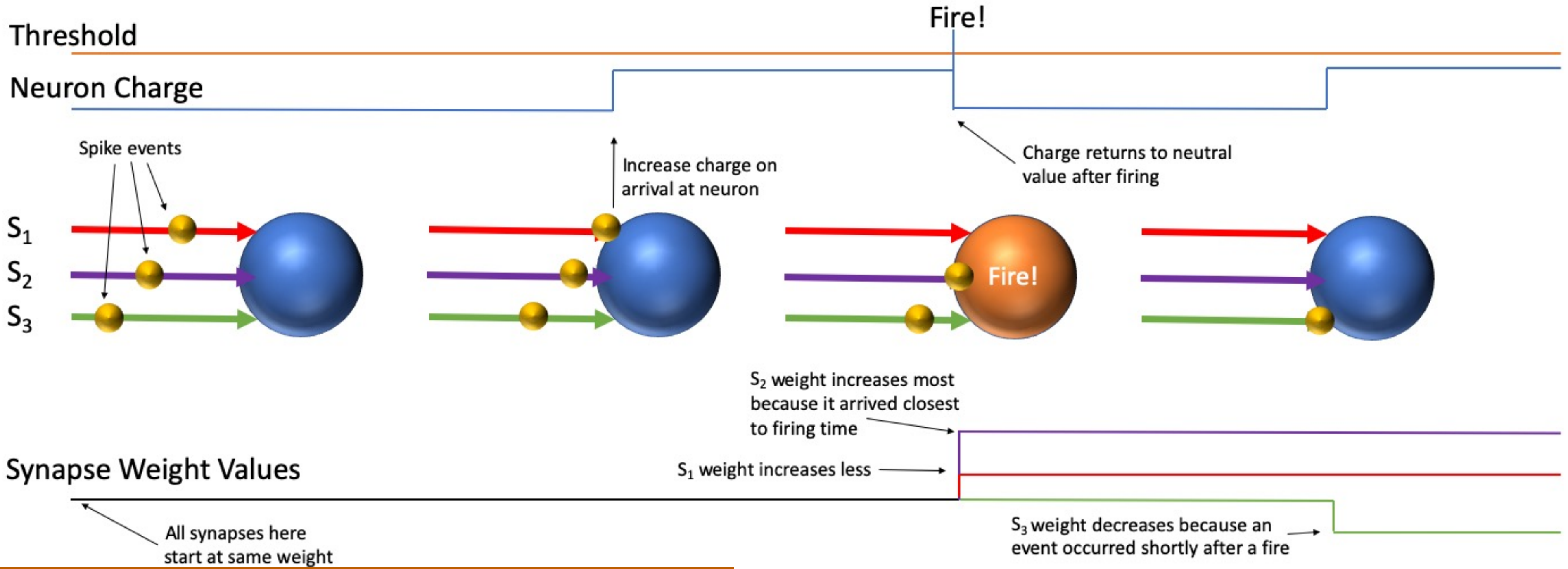
Key Advantage:
Decades of knowledge about traditional ANNs

Key Disadvantage:
Doesn't work natively on many features of SNNs

Algorithms: Synaptic Plasticity

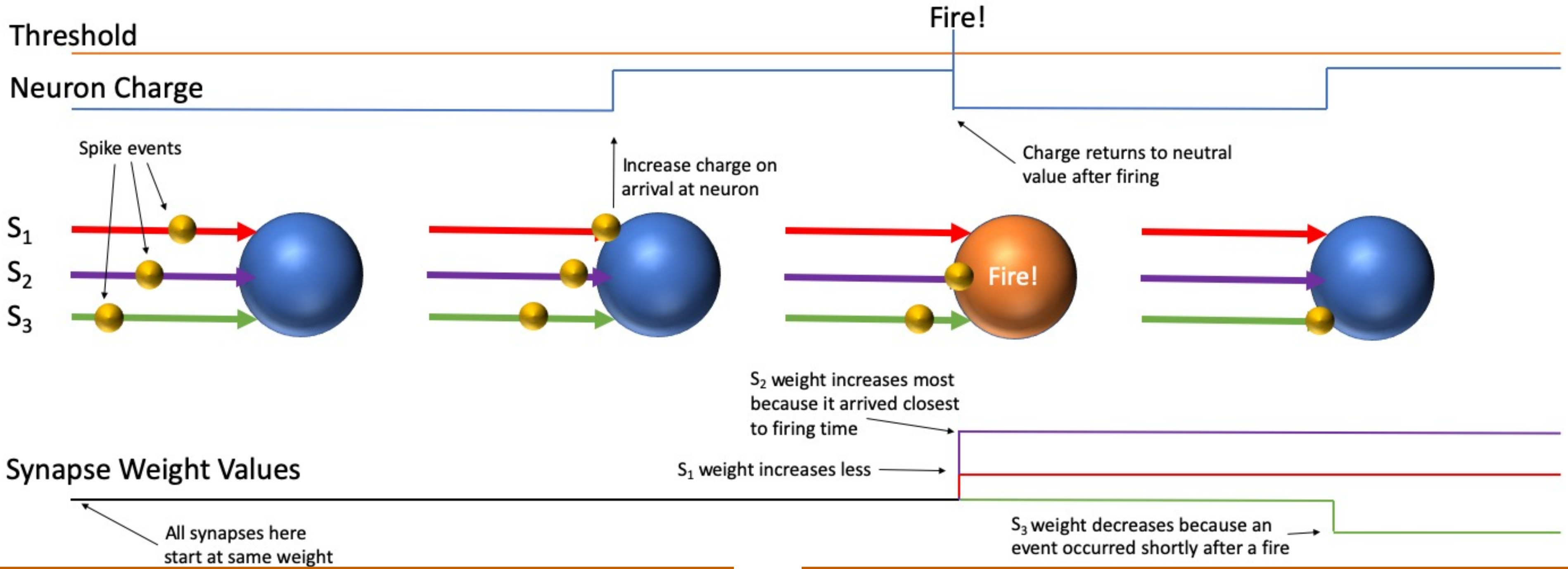


Algorithms: Synaptic Plasticity



Key Advantage: Biologically-inspired and unsupervised

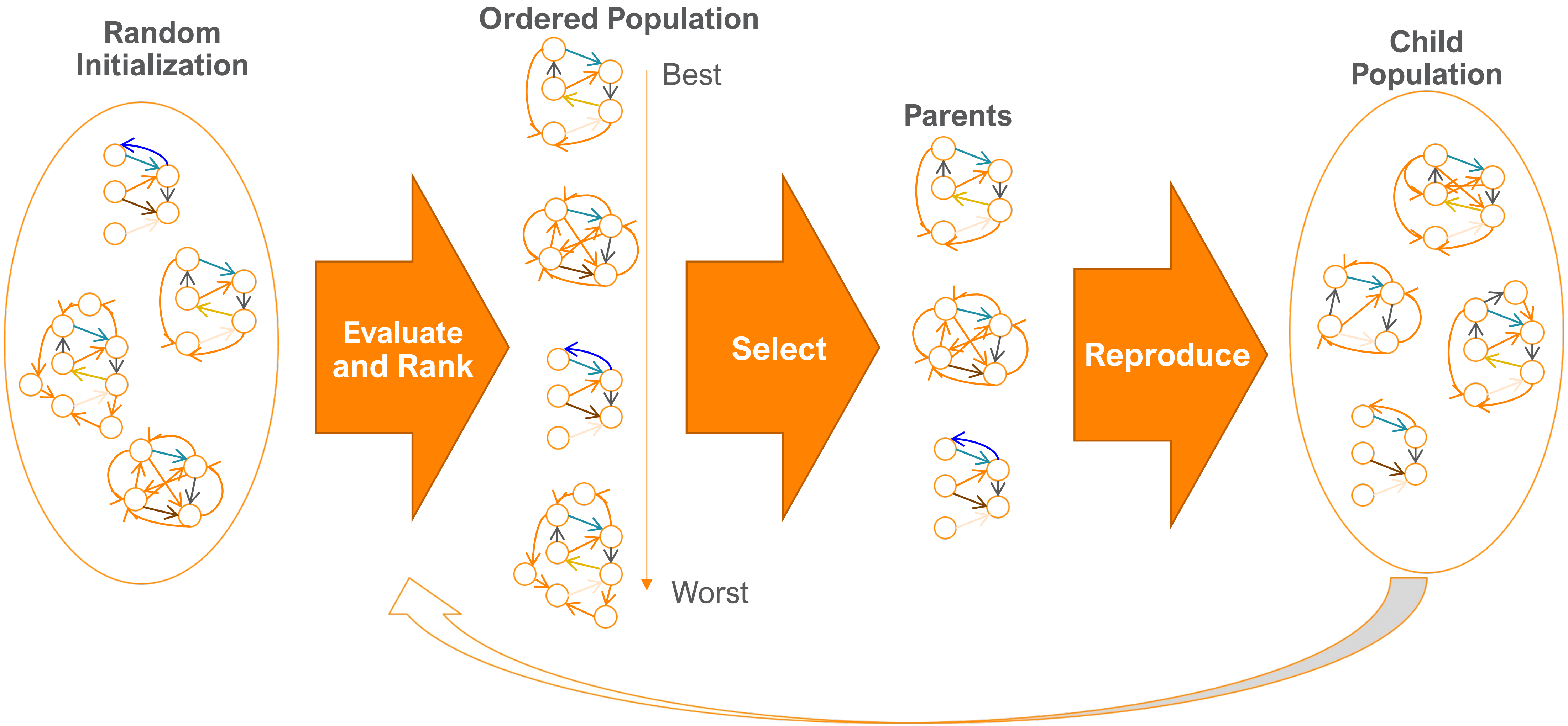
Algorithms: Synaptic Plasticity



Key Advantage: Biologically-inspired and unsupervised

Key Disadvantage: Not well understood and not scalable

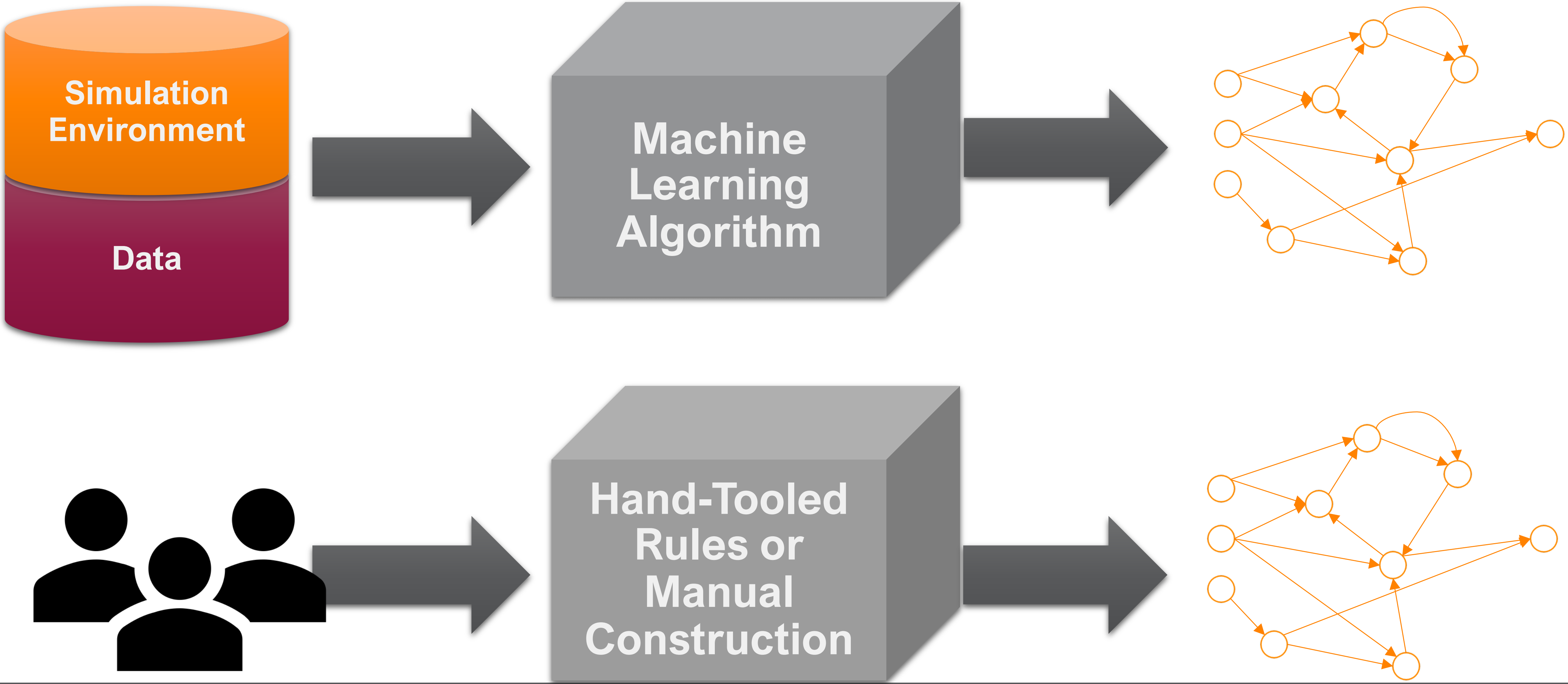
EONS: Evolutionary Optimization for Neuromorphic Systems



Why Evolutionary Optimization?

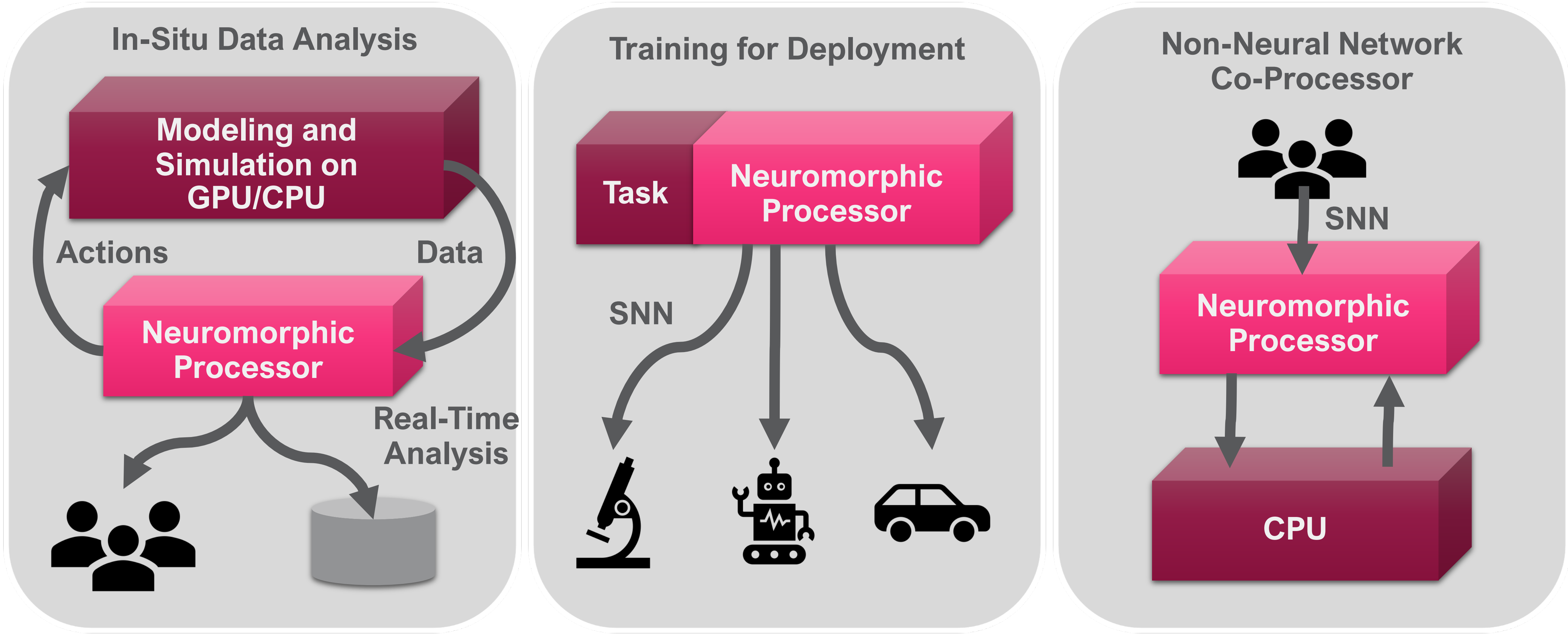
- Applicable to a wide variety of tasks
- Applicable to different architectures and devices
- Operates within the characteristics and constraints of the architecture/device
- Can learn topology and parameters (not just synaptic weights)
- Can interact with software simulations or directly with hardware
- Parallelizable/scalable on HPC

“Programming” via Spiking Neural Networks

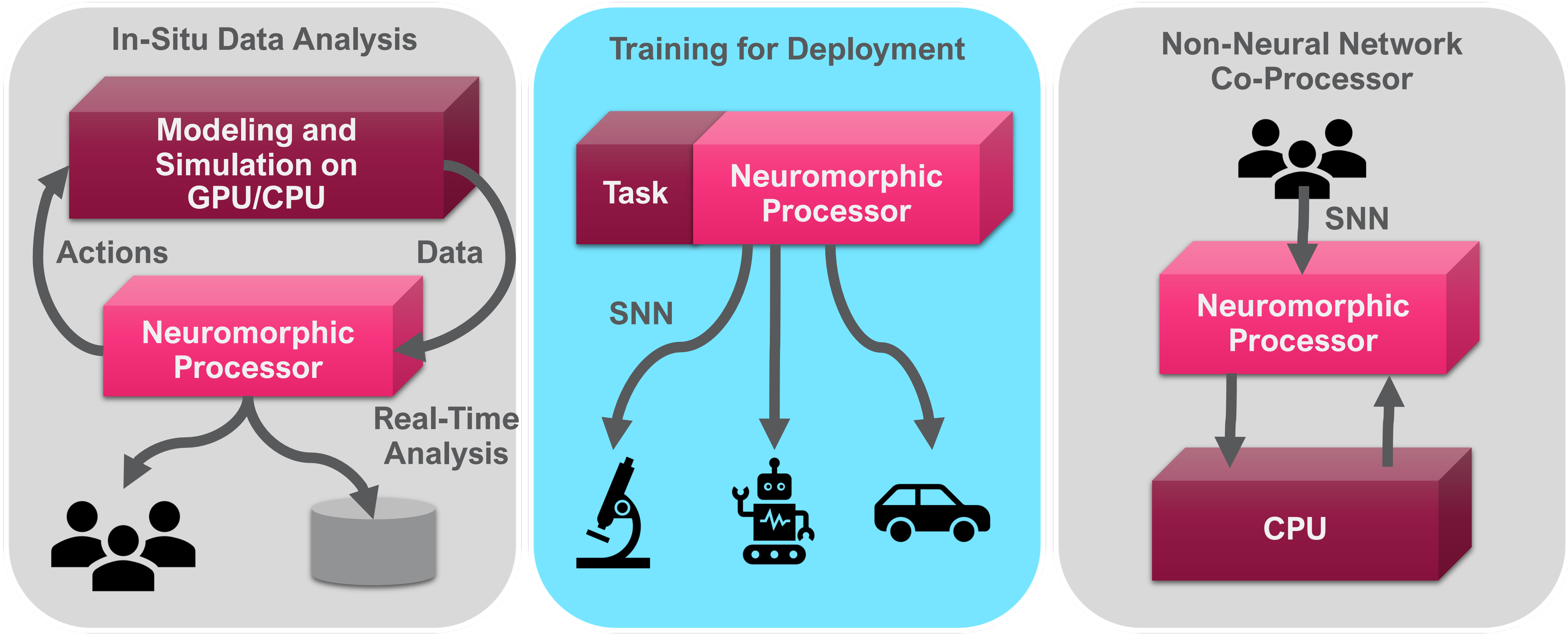


Neuromorphic as a Co-Processor on HPC

Example Neuromorphic Use Cases on HPC

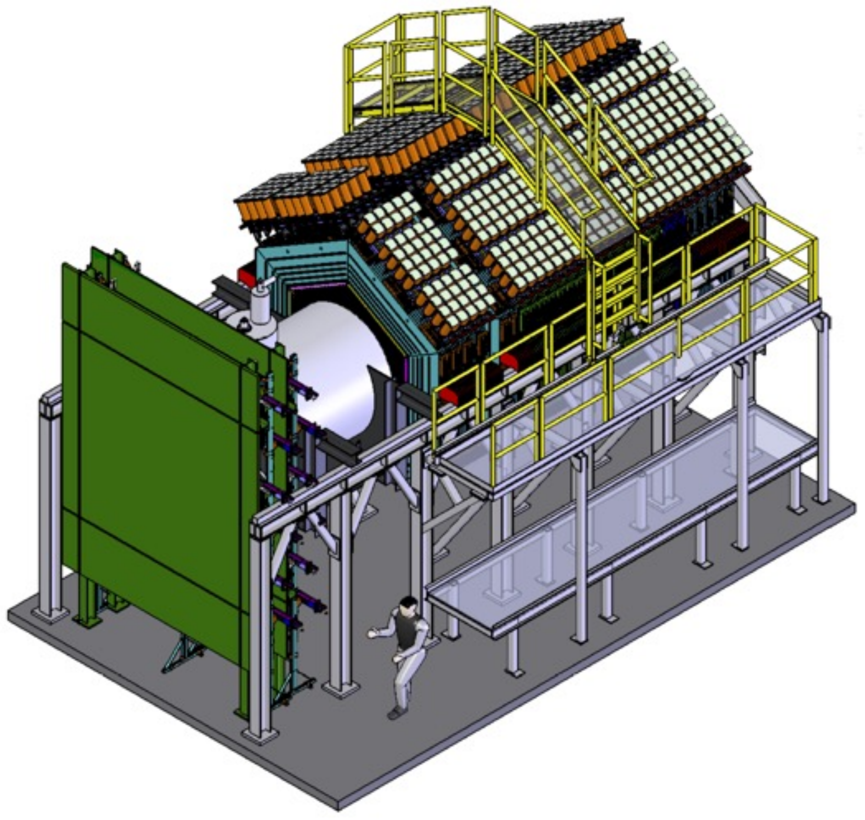


Example Neuromorphic Use Cases on HPC



Data from MINERvA (Main Injector Experiment for v-A)

- Neutrino scattering experiment at Fermi National Accelerator Laboratory
- The detector is exposed to the NuMI (Neutrinos at the Main Injector) neutrino beam
- Millions of simulated neutrino-nucleus scattering events were created
- Classification task is to classify the horizontal region where the interaction originated

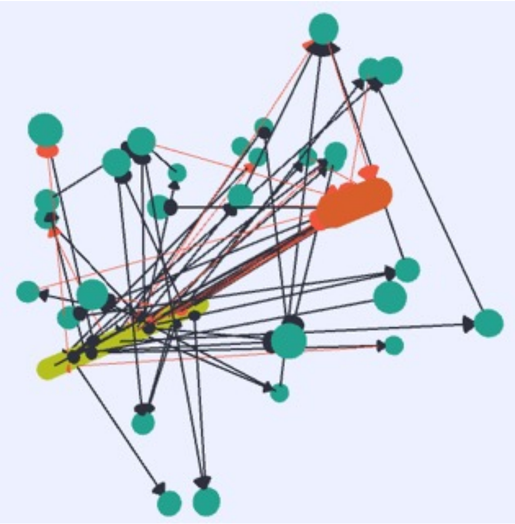


Source: A. Terwilliger, et al. Vertex Reconstruction of Neutrino Interactions using Deep Learning. IJCNN 2017.

Best Results: Single View

x-view (127x50)	conv1 (8x3)	pool1 (2x1)	conv2 (7x3)	pool2 (2x1)	conv3 (6x3)	pool3 (2x1)	conv4 (6x3)	pool4 (2x1)	fc1 (196)	drop out	fc2 (98)	drop out	fc3 (11)	classification

Convolutional Neural Network Result: ~80.42%

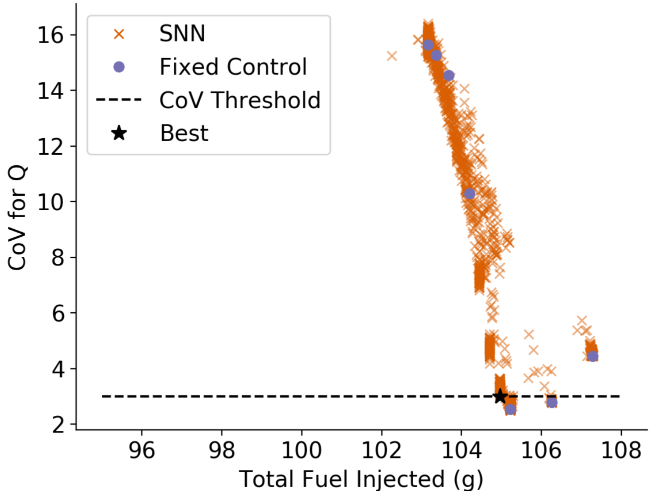


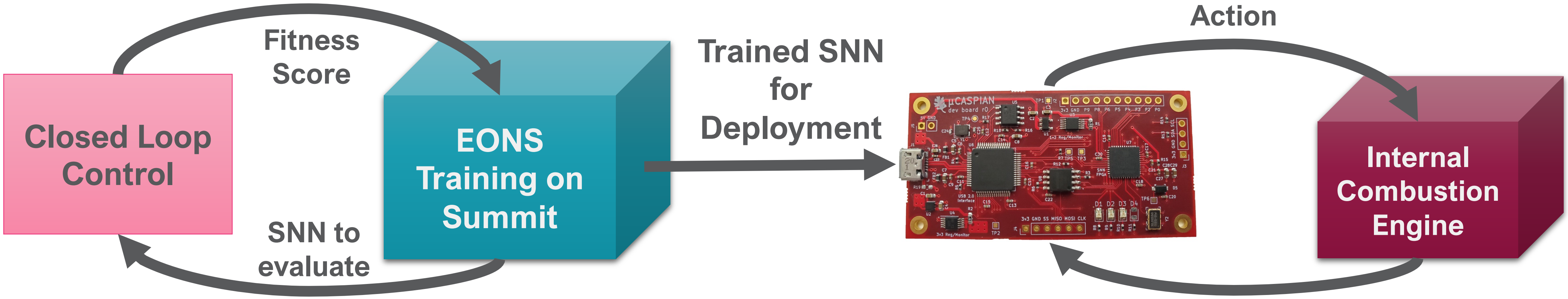
- 90 neurons, 86 synapses
- Estimated energy for a single classification for mrDANNA implementation: 1.66 μJ

Spiking Neural Network Result: ~80.63%

Neuromorphic Engine Control for Fuel Efficiency

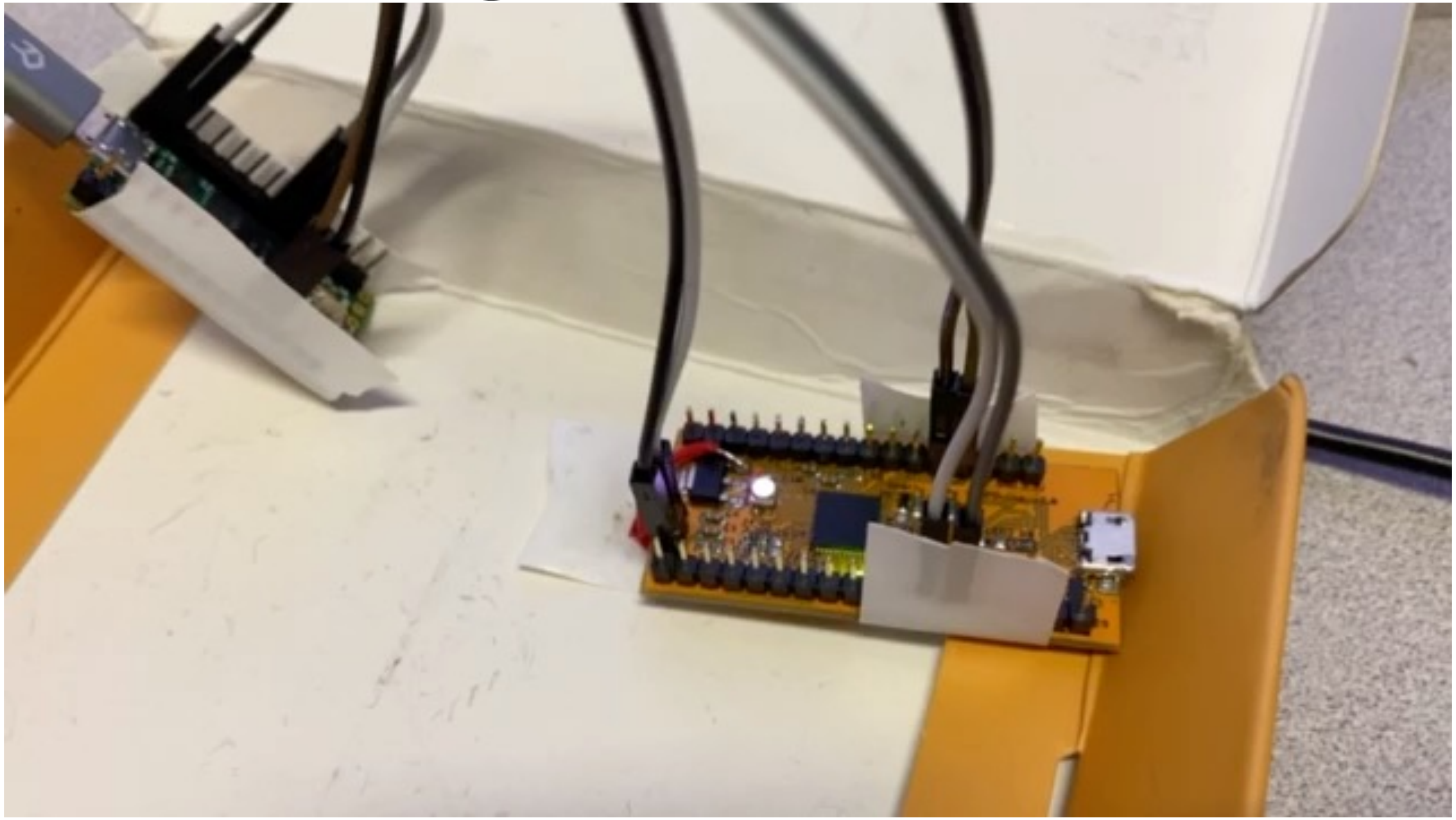
- Developed a complete workflow to train a spiking neural network (SNN) to deploy to an FPGA-based neuromorphic hardware system for internal combustion engine control.
- SNN-based approach outperforms fixed control strategies in terms of fuel efficiency in simulation while still meeting acceptable performance metrics.
- Currently deploying SNN trained on Summit to neuromorphic hardware in-the-loop with engine at National Transportation Research Center.





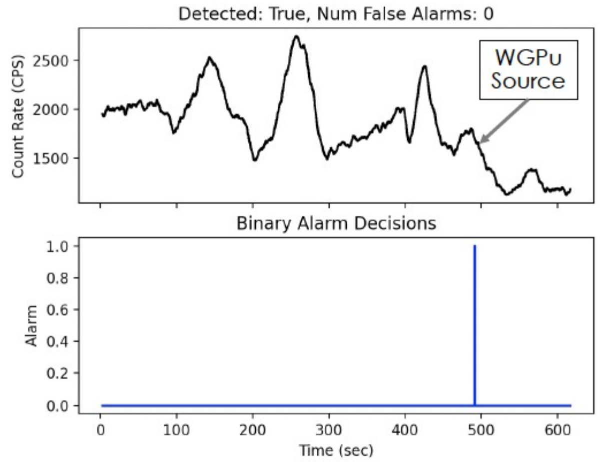
Catherine D. Schuman, Steven R. Young, J. Parker Mitchell, J. Travis Johnston, Derek Rose, Bryan P. Maldonado, Brian C. Kaul. “Low Size, Weight, and Power Neuromorphic Computing to Improve Combustion Engine Efficiency.” International Conference on Green and Sustainable Computing 2020.

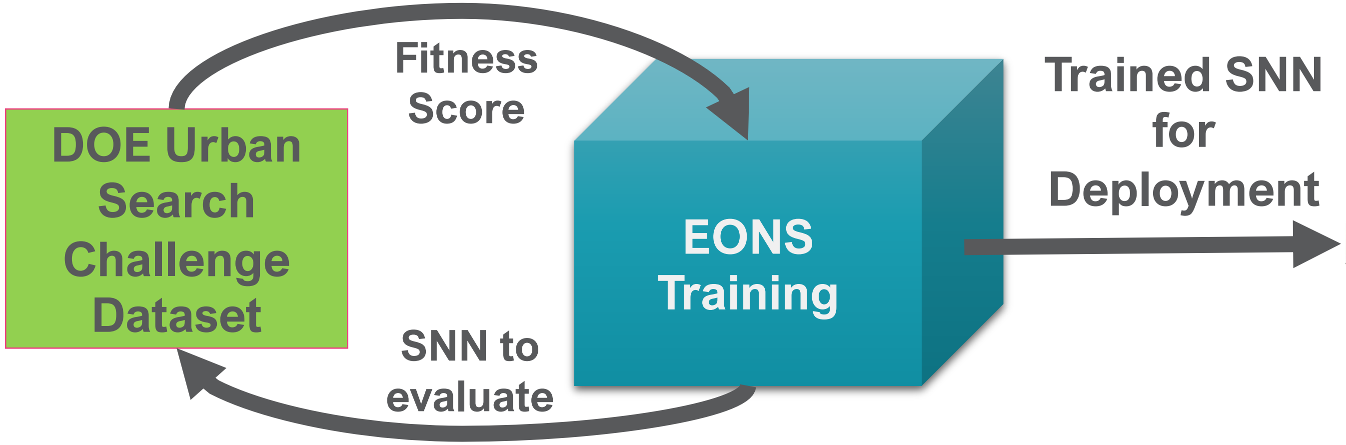
Neuromorphic Engine Control for Fuel Efficiency

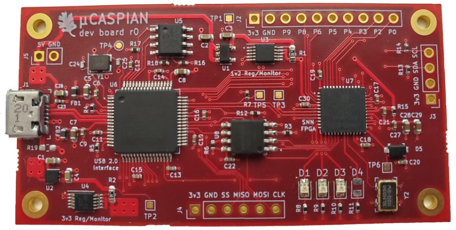


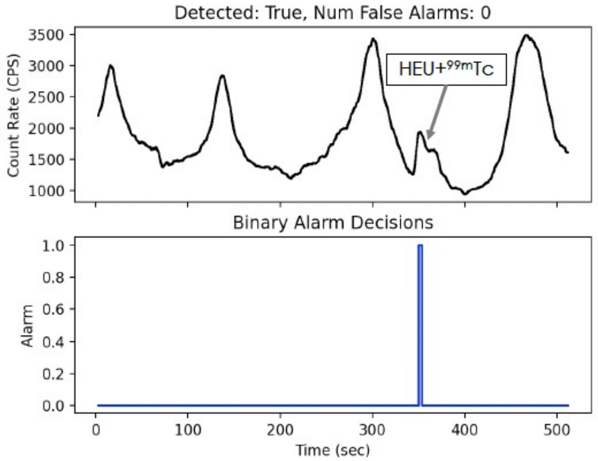
Neuromorphic Radiation Detection

- Radiation detection algorithms must be able to detect low-SNR anomalies in a very noisy and dynamic data environment.
- Neuromorphic computing enables the ability to combine the computational performance of machine learning with massive reductions in power consumption for this task
- K-sigma performance on DOE Urban Search Challenge: F1-Score: 0.080
- Current SNN trained with EONS performance: F1-Score: 0.436





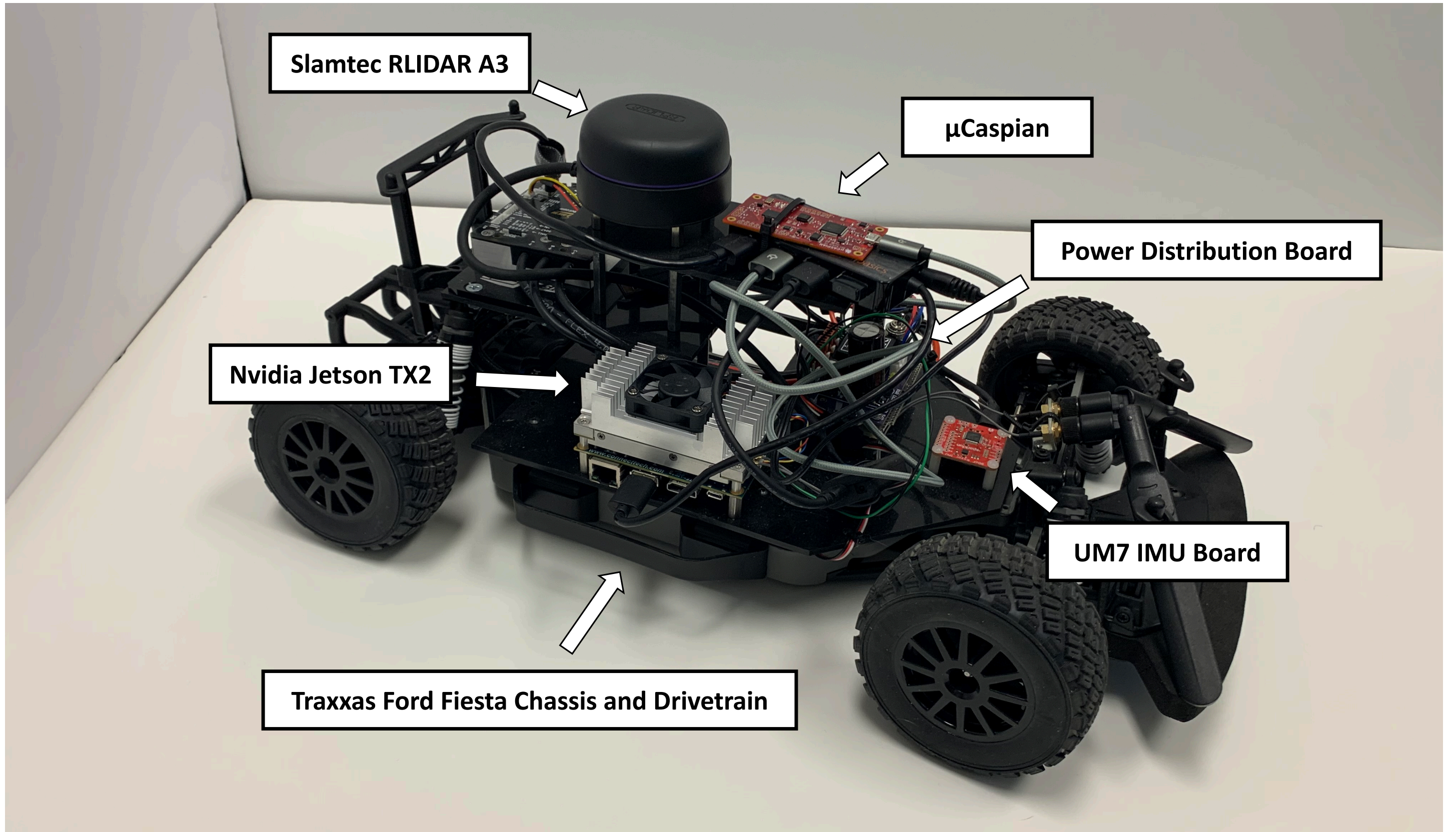




James Ghawaly, Aaron Young, Brett Witherspoon, Dan Archer, Nick Prins and Catherine Schuman. "A Neuromorphic Algorithm for Radiation Anomaly Detection." In preparation.

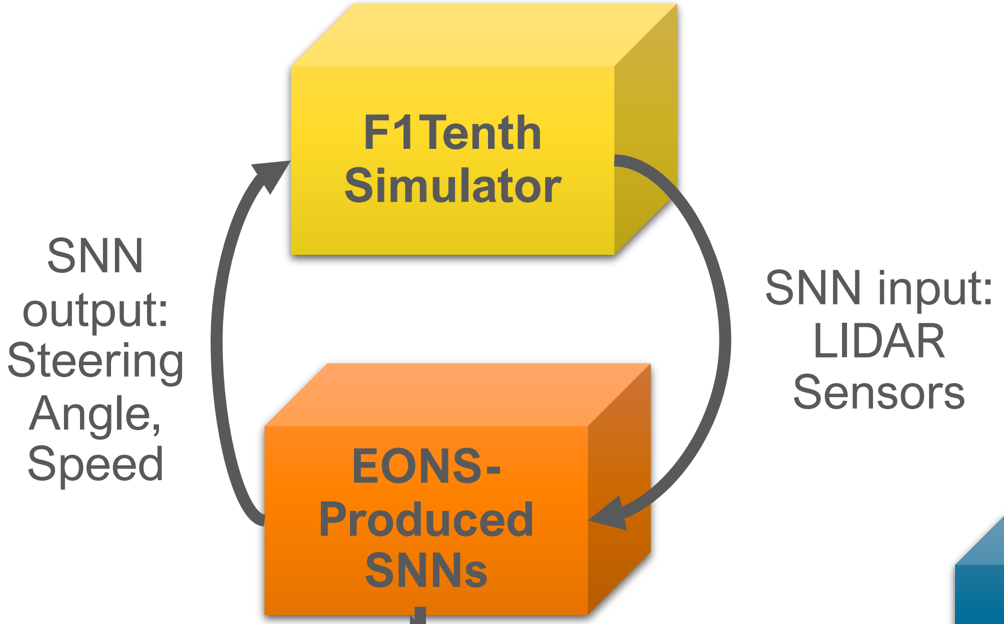
F1Tenth: Autonomous Racing

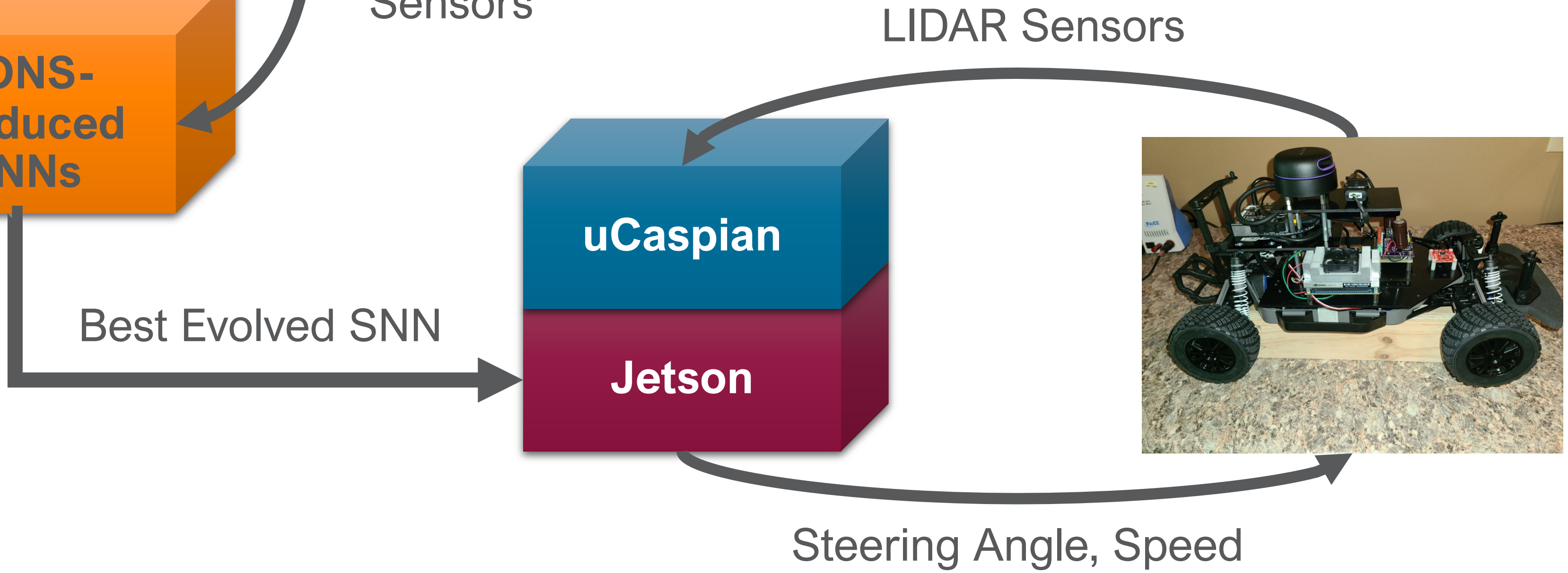
- Fully autonomous 1/10th scale racing of Formula One (https://f1tenth.org/)
- Like full scale vehicles, the need for low size, weight, and power is critical
- Relatively inexpensive real-world demonstration of what neuromorphic computing can provide



F1Tenth: Autonomous Racing

- We use EONS to train SNNs for uCaspian deployment using the F1Tenth Simulator.
- We deployed the trained SNN to a Jetson first (CPU will simulate uCaspian) and then to the actual uCaspian to control the physical car.



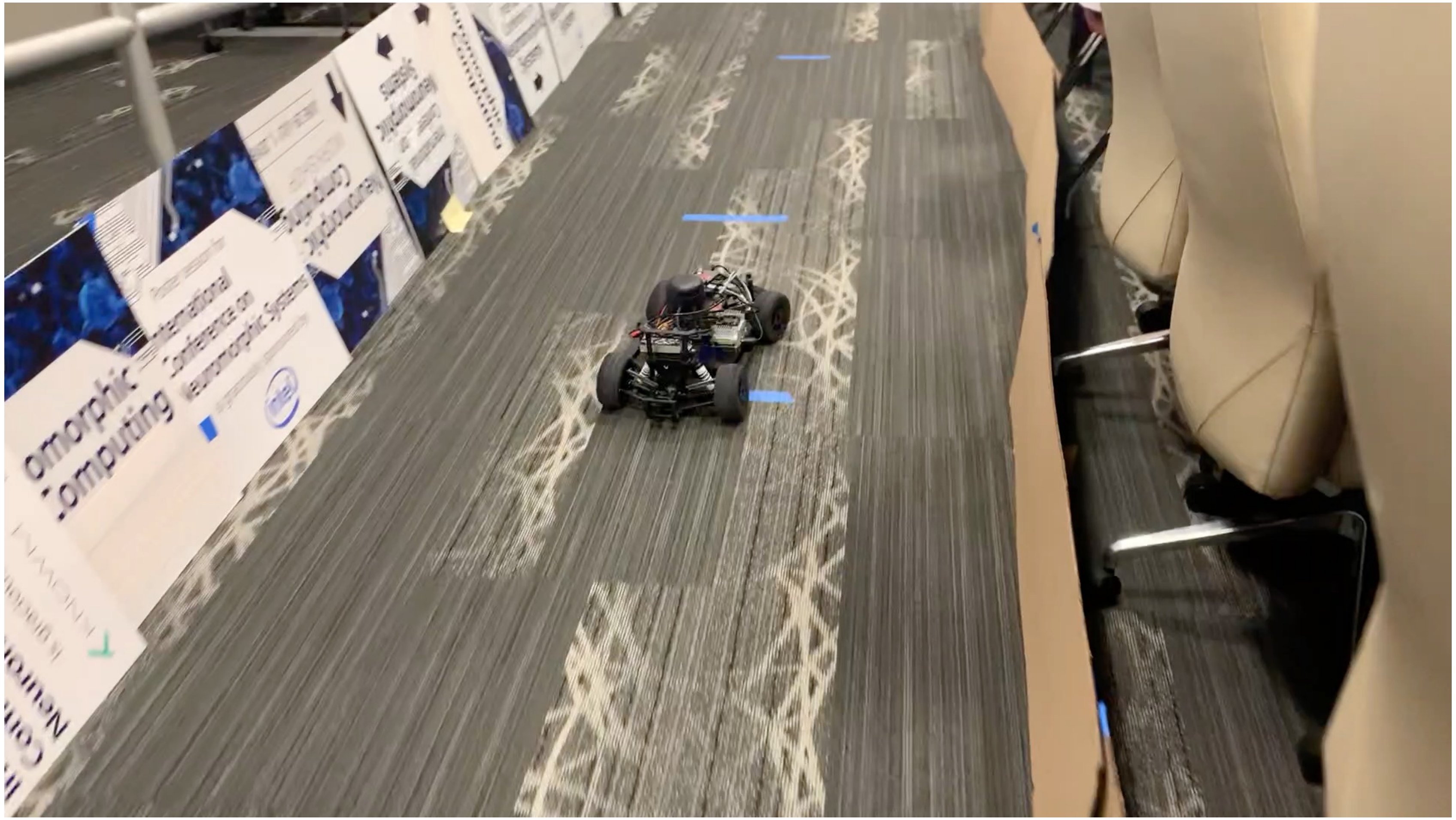


Training Tracks

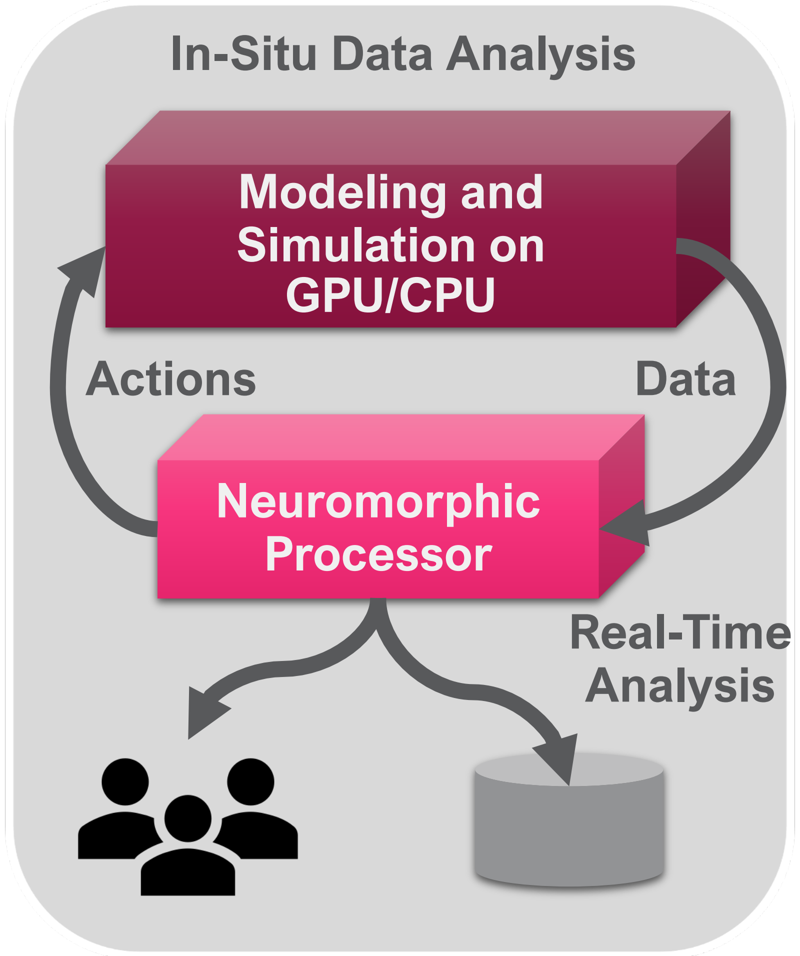
Physical Deployment

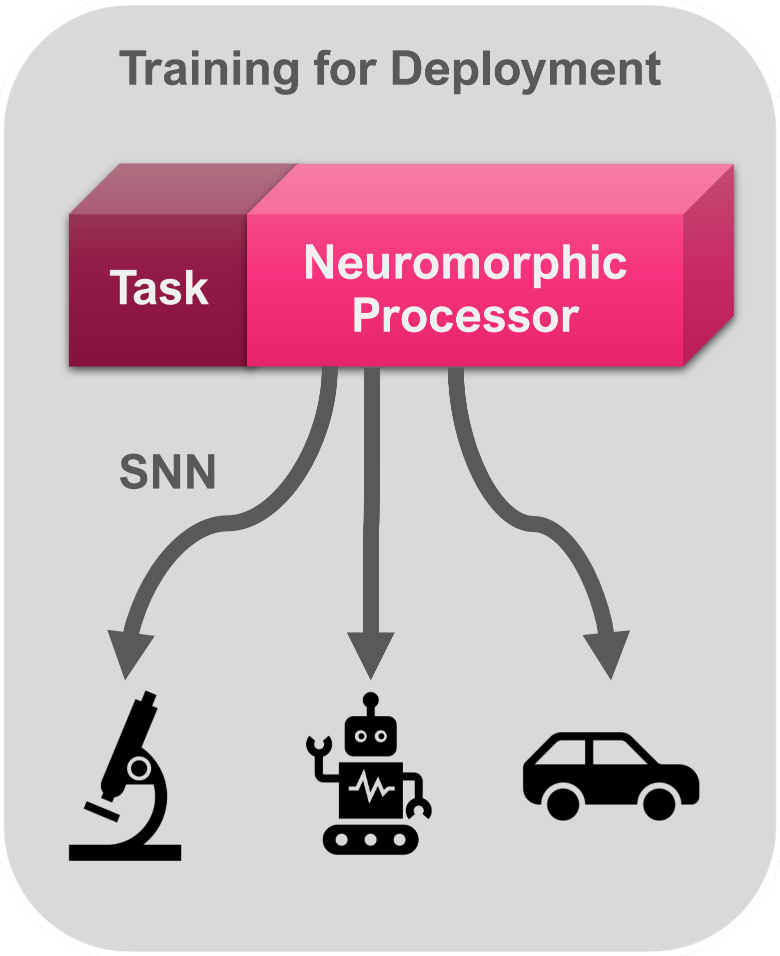
Robert Patton, Catherine Schuman, Shruti R. Kulkarni, John Mitchell, N. Quentin Haas, Christopher Stahl, Spencer Paulissen, Prasanna Date, Thomas Potok, Shay Snyder and Maryam Parsa, “Neuromorphic Computing for Autonomous Racing.” International Conference on Neuromorphic Systems (ICONS) 2021.

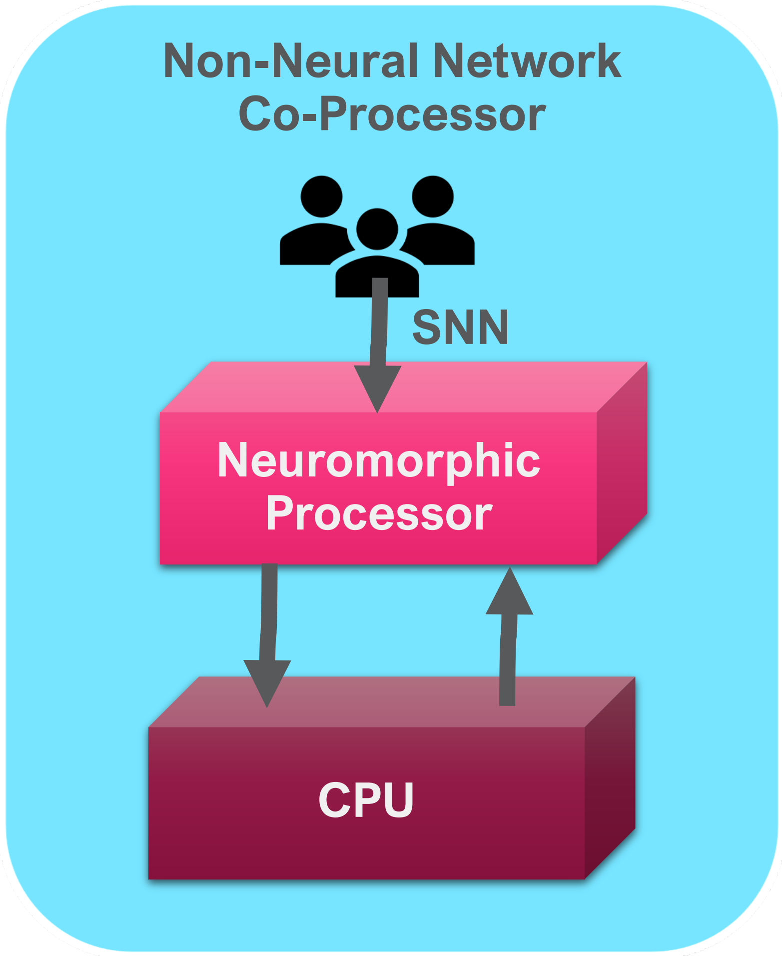
Physical Deployment



Example Neuromorphic Use Cases on HPC







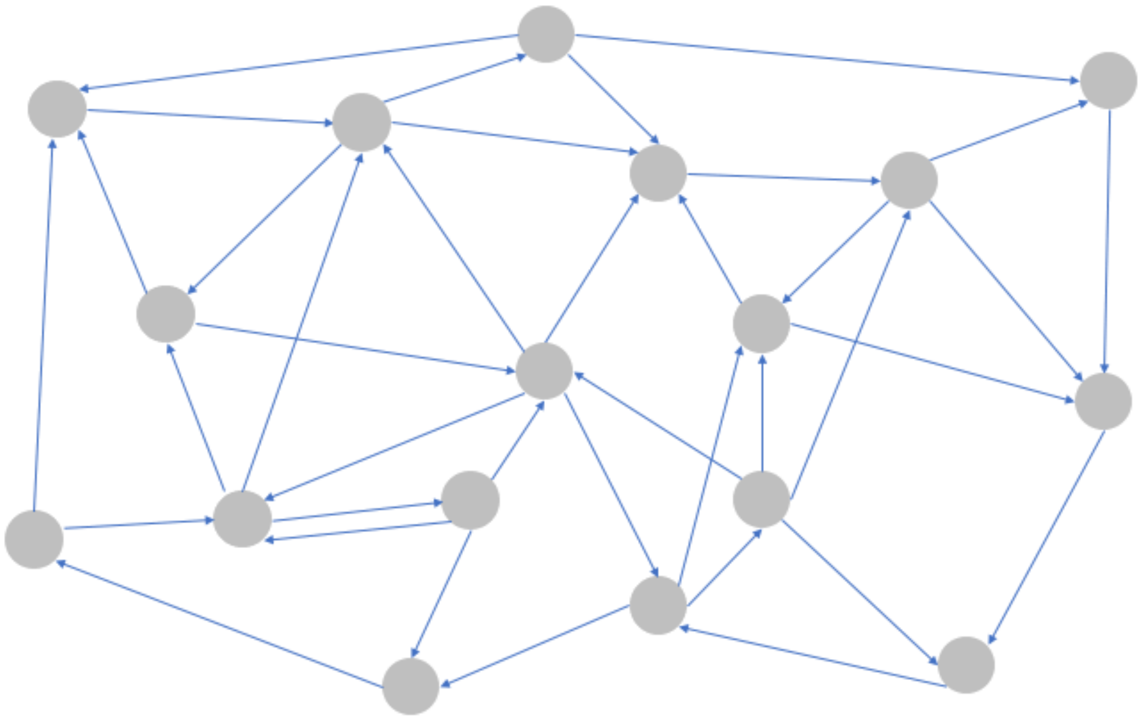
Properties of Spiking Neuromorphic Systems

- Massively parallel computation
- Collocated processing and memory
- Simple processing elements that perform specific computations
- Simple communication between elements
- Event driven computation
- Stochastically firing neurons for noise
- Inherently scalable architectures

These properties are useful for more than just machine learning algorithms!

Calculating Shortest Paths

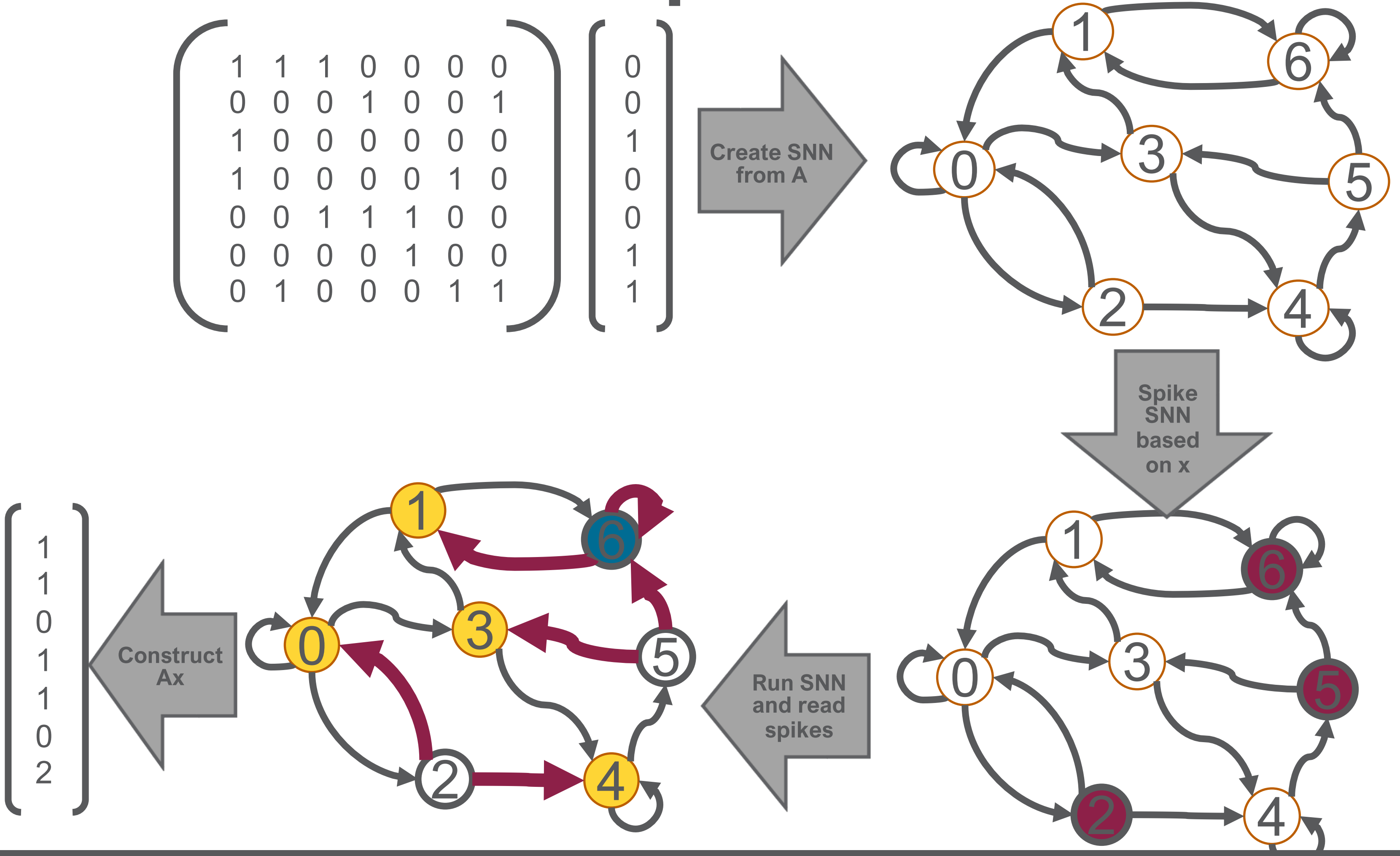
- Graphs are converted into networks
- Distances are converted to delays
- Spikes travel throughout the network and give single-source shortest path lengths



Schuman, Catherine D., Kathleen Hamilton, Tiffany Mintz, Md Musabbir Adnan, Bon Woong Ku, Sung-Kyu Lim, and Garrett S. Rose. "Shortest path and neighborhood subgraph extraction on a spiking memristive neuromorphic implementation." In *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop*, pp. 1-6. 2019.

Sparse Binary Matrix-Vector Multiplication

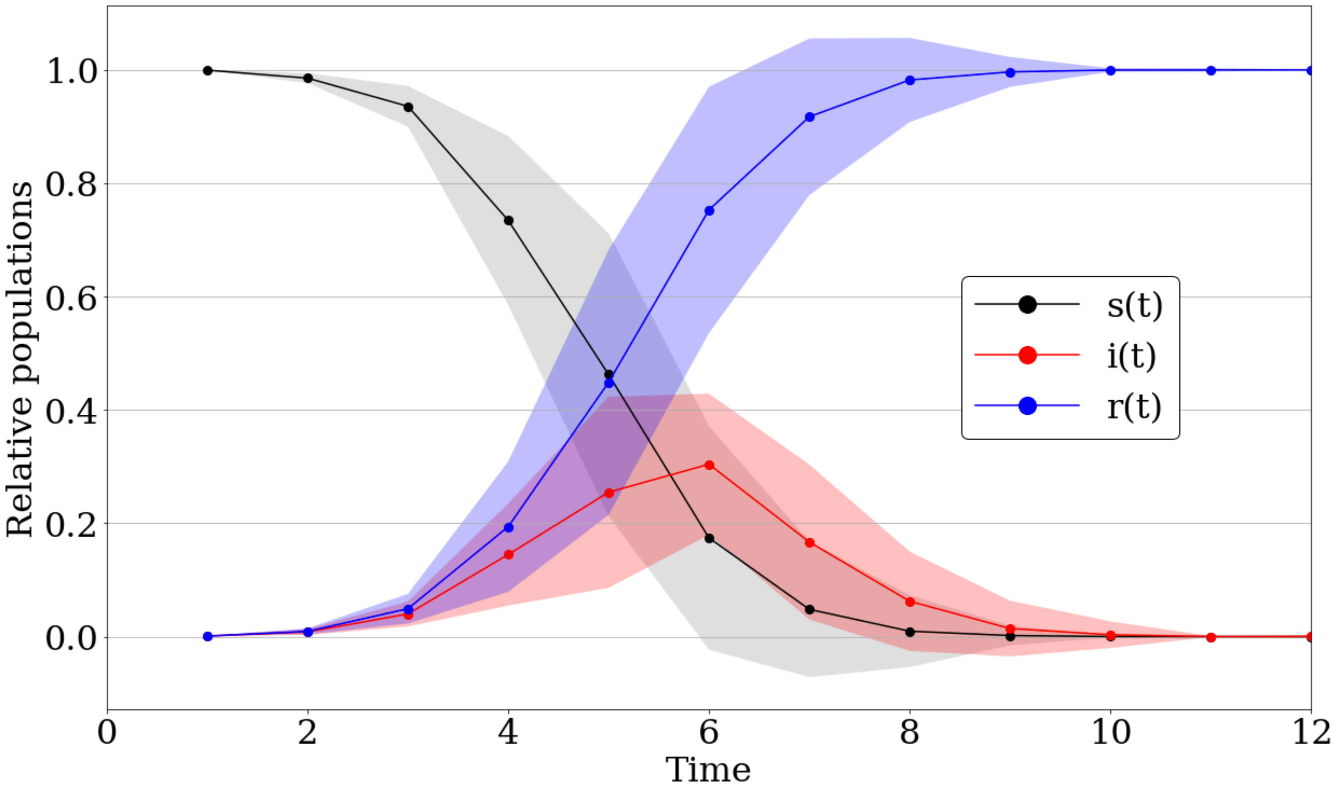
- We demonstrated that binary matrix-vector multiplication can be computed using networks of spiking neurons
- Next steps: Evaluate on real neuromorphic hardware



Catherine D. Schuman, Bill Kay, Prasanna Date, Ramakrishnan Kannan, Piyush Sao, and Thomas E. Potok. "Sparse Binary Matrix-Vector Multiplication on Neuromorphic Computers." GrAPL 2021: Workshop on Graphs, Architectures, Programming, and Learning,

Modeling Epidemic Spread

- Neurons are individuals in a population
- Synapses are shared social connections
- Spikes are transmission of infection
- Parameters allow for different conditions

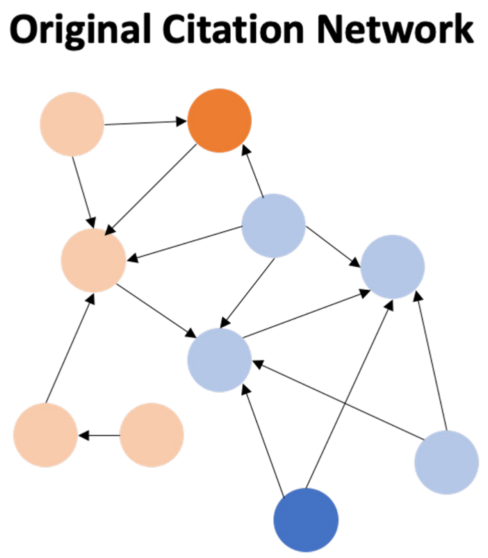


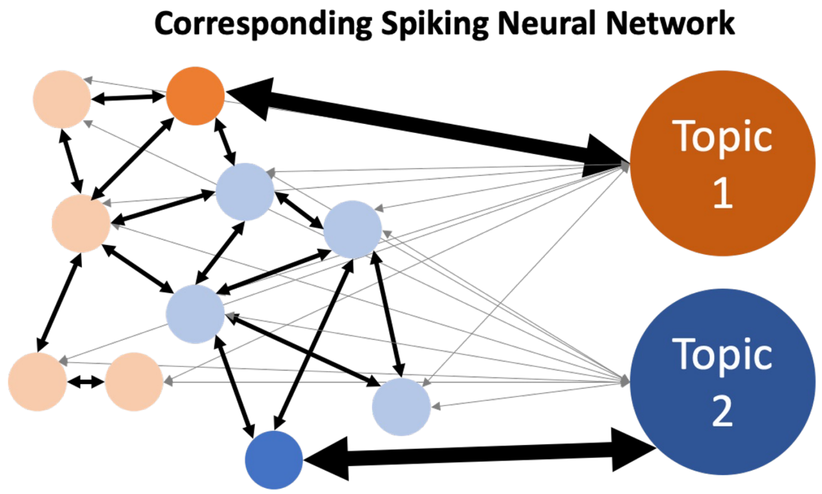
Kathleen Hamilton, Prasanna Date, Bill Kay, and Catherine Schuman D. "Modeling epidemic spread with spike-based models." In *International Conference on Neuromorphic Systems 2020*, pp. 1-5. 2020.

Graph Neural Networks

- Node classification task, without features
  - Citation networks as benchmark datasets for GNNs

	Cora	Citeseer	Pubmed
Node2Vec	0.71	0.48	0.70
Node2Vec-a	0.68	0.51	-
Planetoid-G	0.69	0.49	0.66
GraphSAGE	0.71	0.48	0.64
GCN	0.59	0.34	0.42
Neuromorphic	0.67	0.51	0.79





Guojing Cong, Seung-Hwan Lim, Shruti Kulkarni, Prasanna Date, Thomas Potok, Shay Snyder, Maryam Parsa, and Catherine Schuman. 2018. Semi-Supervised Graph Structure Learning on Neuromorphic Computers. In Proceedings of International Conference on Neuromorphic Systems (ICONS '22). ACM, New York, NY, US

Summary

- Neuromorphic computers are a new type of computer inspired by biological brains
- They are “programmed” using spiking neural networks, a more biologically inspired neural network
- We have successfully applied neuromorphic to a wide variety of applications, including scientific data analysis and robotics
- Neuromorphic computers are useful for more than just neural network computation!

Looking for postdoc or graduate opportunities?

TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

- We're actively recruiting Master's and PhD students across the TENNLab research group
- We are also recruiting postdocs specifically in software, algorithms, and application development


Dr. Ahmed Aziz


Dr. Garrett Rose


Dr. Jim Plank


Dr. Katie Schuman

TENN LAB
NEUROMORPHIC
ARCHITECTURES. LEARNING. APPLICATIONS.

Work supported by:

Department of Energy

Air Force Research Lab

OAK RIDGE
National Laboratory












Thank you!

Questions?

Contact:
Email: cschuman@utk.edu

Website: catherineschuman.com

Twitter: @cdschuman